

AD#642829

ESD ACCESSION LIST

ESD-TR-66-405

ESTI Call No. AI 53659

Copy No. 1 of 2 cys.



STUDY AND TEST OF A METHODOLOGY FOR LABORATORY EVALUATION OF MESSAGE RETRIEVAL SYSTEMS

Vincent E. Giuliano

Paul E. Jones

ESD RECORD COPY

RETURN TO
SCIENTIFIC & TECHNICAL INFORMATION DIVISION
(ESTI), BUILDING 1211

August 1966

ARTHUR D. LITTLE INC.

Interim Report

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Mass.

Distribution of this document
is unlimited.

(Prepared under Contract No. AF 19(628)-4067 by
Arthur D. Little, Incorporated, Cambridge, Mass.)

AD0642829

When U. S. Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Do not return this copy. Retain or destroy.

STUDY AND TEST OF A METHODOLOGY FOR
LABORATORY EVALUATION OF MESSAGE RETRIEVAL SYSTEMS

Vincent E. Giuliano

Paul E. Jones

August 1966

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

C-65850
C-66257

(Prepared under Contract Nos. AF 19(628)-4067 and 3311
by Arthur D. Little, Incorporated, Cambridge, Mass.)

FOREWORD

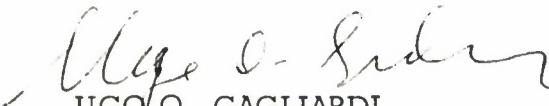
This work was conducted in support of Project 2806, Task 280601 by Arthur D. Little, Inc., 35 Acorn Park, Cambridge, Mass. under Contract AF 19 (628) - 4067. Our internal code for this contract is C-66257. The work was also supported in part under Contract AF 19 (628) - 3311.

The program was monitored for the U. S. Air Force by John B. Goodenough ESRHT.

Work reported herein was performed during the period March 1964 to July 1966, and the draft report was submitted on 14 July 1966. We wish to acknowledge the important contribution of Philip Hankins, Inc. in performing part of the computer programming under subcontract to us.

The cooperation between the Science and Technology Information Division of NASA and the Decision Sciences Laboratory has been of immense value to the research reported here. In particular, programs developed under our Contract NASW-1051 with NASA have been used in the investigations reported here, and conversely.

This technical report has been reviewed and is approved.


UGO O. GAGLIARDI
Decision Techniques Division
Decision Sciences Laboratory

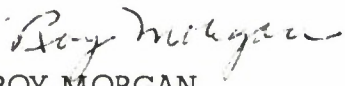

ROY MORGAN
Colonel, USAF
Director, Decision Sciences Laboratory
Deputy for Command Systems

TABLE OF CONTENTS

	<u>Page</u>
List of Tables	vii
Abstract	ix
I. INTRODUCTION	1
A. Purpose and Scope	1
B. Method	5
C. General Findings About Message Retrieval	6
D. Specific Findings Relating to Message Retrieval .	11
E. General Observations Relating to Retrieval Methodology	14
F. Specific Observations Relating to Evaluation Methodology	16
II. THE MESSAGE RETRIEVAL EVALUATION PROBLEM	19
A. Identification of Retrieval Parameters	19
B. Approaches to Evaluation	22
C. Concerning Relevance	26
D. Previous Work	28
III. EXPERIMENTAL DESIGN CONSIDERATIONS	30
A. Collection Size and Sampling Strategies	30
B. Selecting Queries	32
C. Making and Recording Relevance Judgments	36
D. Economics of Evaluation	40
E. Insight-Oriented vs. Proof-Oriented Tests	43

TABLE OF CONTENTS (Continued)

	<u>Page</u>
IV. MEASURES AND TOOLS FOR EVALUATION AND COMPARISON OF INFORMATION RETRIEVAL SYSTEMS	45
A. Introduction	45
B. Definitions	45
C. Previous Statistics and Measures	48
D. The Disadvantages of Previous Statistics	54
E. Toward a Rational Measure of Effectiveness	56
F. Performance Characteristic Curves	57
G. Measures of Performance Features	60
H. Modeling Retrieval System Performance	62
I. Summary	63
V. EXPERIMENTAL DATA BASES AND RETRIEVAL TOOLS	64
A. Data Bases	64
B. Comparison of Indexing Vocabularies and Message Indexing Coverage	70
C. Retrieval Searching Tools	82
VI. EXPERIMENTAL RESULTS	89
A. Retrieving on Subject-Heading Type Queries	89
B. Retrieving on Full-Text Queries	115
C. Multi-Evaluator Tests	126
D. Manual vs. Automatic Indexing	128
REFERENCES	135

TABLE OF CONTENTS (Continued)

	<u>Page</u>
APPENDIX A ESTABLISHING A MASTER LIST: COMBINING THE JUDGMENTS OF SEVERAL EVALUATIONS	139
APPENDIX B MEASURES OF USER SATISFACTION WHICH RELATE TO SEARCH OBJECTIVES	145
APPENDIX C EXTRAPOLATION OF OVERLAP RESULTS AND ESTIMATES OF COLLECTION PARAMETERS	161
APPENDIX D RELIABILITY TEST OF NASA - GE OVERLAP PROPORTIONS	171

LIST OF TABLES

<u>Table No.</u>		<u>Page</u>
IV-1	Performance Characteristic Curves For Three Hypothetical Search Options	58
IV-2	Actual Performance Characteristic Curve for the Query "Surface Strain"	59
V-1	Sample Abstracts From the GE-2 Corpus	66
V-2	Message "Length" Distribution for GE-2A Data.	67
V-3	Some Statistics of GE-2 Corpus Used for Associative Retrieval Experiments	69
V-4	GE-2A Terms Included in the UNITERM Vocabulary	72
V-5	Morphological Cognates	73
V-6	GE-2A Test Words Without UNITERM Cognates	74
V-7	Occurrence Frequency Correlation of Terms Common to GE-1 and GE-2 Vocabularies	75
V-8	Example Worksheet for Message #59512	78
V-9A	Distribution of Spurious Term Set Size.	81
V-9B	Comparison of Measures for Entire Sample and Subpopulations R and Q	81
V-10	Initial Portions of Four Typical Association Lists \sqrt{A} KA Matrix, GE-2 Vocabulary	83
V-11A	Typical Input Query to the Phase I Program	85
V-11B	Association Profile Produced by the Phase I Program for the Input Query of Table V-11A	86
V-12	Examples of Content-Bearing Pairs Retrieved by the Phase III Program	88
VI-1	Precision of Search Plotted Against C_{ab} Value	99
VI-2	Association List Generated for the Query "Surface Strain"	107

LIST OF TABLES (continued)

<u>Table No.</u>		<u>Page</u>
VI-3	Performance Characteristic Curves for the Subject-Heading Query "Surface Strain"	109
VI-4	Comparison of "Conventional Coordinate" and "Automatic Associative" Retrieval for Queries which are System CBUs -- Average Results for Twelve Queries	110
VI-5	GE-2A Vocabulary Coverage.	117
VI-6A	Original Query	118
VI-6B	Interrogative Version of Same Query	118
VI-7	Average Performance Characteristic Curves for the Six Types of Requests Over the Four Questions	122
VI-8	Estimated Fractions of Relevance Points	125

ABSTRACT

This report documents two years of work on the laboratory evaluation of message and document retrieval systems. It contains a general discussion of the problems of laboratory evaluation of retrieval systems, and specific findings relating both to the methodology of evaluation and search performance results observed with a large-scale experimental collection.

The initial sections of the report are devoted to developing a general framework for viewing the problems of performance evaluation under laboratory conditions. We identify and discuss several mathematical techniques potentially useful in the evaluation process, including methods for unbiasing and averaging the results of judgments by several independent evaluators. Also, many possible measures of system performance are discussed, compared, and evaluated.

We describe the processing of our 10,000-message experimental collection, including the steps of automatic indexing and computation of word-association measures. Comparison of subject matter coverage and effects of manual and automatic indexing for this collection are discussed, and several statistical characterizations of our collection are presented.

We also describe several experimental forays with our collection using combinations of conventional and associative retrieval with and without human intervention, using multiple evaluators, and we consider both full text and subject heading queries.

Numerous conclusions and findings are presented with respect to efficacy of various retrieval evaluation techniques and methods, the relative merits of machine and automatic indexing, and the comparative efficacy of various combinations of conventional and associative search options.

The report should be of general interest to all concerned with advanced library documentation systems and automatic language processing methods, and of specific interest to workers concerned with evaluating information storage and retrieval systems.

LABORATORY EVALUATION OF MESSAGE RETRIEVAL SYSTEMS

SECTION I INTRODUCTION

A. Purpose and Scope

This interim report treats our work on laboratory evaluation of associative message retrieval methods. The main objectives of our analyses include:

- * A research study of techniques for the evaluation of associative and coordinate message retrieval methods.
- * The preparation and analysis of machine-readable test corpora suitable for experiments with associative message retrieval methods. These corpora consist of approximately 10,000 messages and 1,000 index terms each; one is manually indexed, the other is automatically indexed.
- * Experimental application of the evaluation techniques to the test corpora with the aim of achieving a comparison between conventional and associative retrieval methods.
- * Investigation of the possibility of extending the associative retrieval methods to apply to very large data bases in conjunction with real-time processing modes.

In addition to these explicit objectives, we have worked towards several additional goals. Having constructed a prototype associative retrieval system, a key objective was to learn more about the search strategies which are effective in conjunction with associative retrieval on large machine-indexed collections. In particular, we wanted to know whether human mediation in the associative search process is beneficial, and we wanted an idea of what the mediation should consist. Another objective was to identify more clearly the types of information requests for which associative retrieval is significantly better than conventional methods and to obtain numerical estimates of how much better. We wanted to learn something about performance evaluation measures and determine whether any of the many which have been proposed are particularly appropriate for the message retrieval evaluation application.

Another significant objective was to obtain insight as to whether additional techniques of automatic language-processing -- beyond the simple framework of automatic indexing by words, computation and use

of association matrices -- would give further improvements in the message retrieval process. A final objective was to learn how to structure the problem of evaluating large-scale associative message retrieval systems so that we could be prepared, as a next step, to evaluate the functioning of such systems within operational contexts.

1. Major Departures

The prototype information retrieval system we have studied incorporates three major departures from the automatic searching frameworks upon which the bulk of formal evaluative effort has been directed in the recent literature. Unlike those systems which employ the logical/coordinate-retrieval methodology, the prototype produces a ranked output of retrieved messages (documents) rather than selecting a set of items. Second, because a multiplicity of opportunities are available for making use of the association profiles during search, the prototype cannot be characterized as a single fixed system. It is more accurate to see it as several systems, some of which employ human interaction during search and some of which do not. Finally, the prototype system incorporates automatic indexing of the message texts.

As a consequence of these three departures, our laboratory study of the system's performance attributes has emphasized corresponding aspects of the evaluation problem.

Because the prototype produces ranked output, the task of measuring system performance is somewhat more complicated: familiar summary figures like the precision ratio assume dichotomies (retrieved - not retrieved; relevant - not relevant), but systems which arrange documents in a ranked order of presentation do not conform to these assumptions. Since these and similar dichotomous summary figures do not directly reveal whether the ordering is useful, an investigation was performed to find ways of detecting and measuring the extent to which a given arrangement of documents meets the requestor's search objectives.

The second departure, i.e., the need to deal with the existence of many alternative ways for making use of the term associations during searches, has the principal effect of magnifying all the standard problems of evaluating any retrieval system. Appraising which of the various alternatives are most useful is a vital step in the laboratory study of a prototype; in our efforts to explore the system's behavior, we conducted the same search in a variety of distinguishable ways, and faced the task of "comparing" the quality of the results. It is fairly clear that the problems involved in comparing numerous alternatives in the same system are not much different from the problems of evaluating/ comparing a whole set of different retrieval systems simultaneously. To provide a basis for doing the comparisons objectively, we pursued the development of general frameworks for comparing many outputs.

A significant complicating factor we have had to consider is that performance criteria developed for single-step retrieval processing may be quite inappropriate for multiple-step interactive searching. If it is easy for the person guiding the search to exploit the association clues to "home in" on desired material, it may be unimportant whether the measured performance of any given small step in isolation is good or poor -- the crucial question is the effectiveness of the over-all iterative process.

Finally, the third feature of the prototype that encompasses less familiar territory is the presence of an automatically indexed collection. Data on automatic indexing effectiveness is sparse, and a study of the automatic indexing alone (e.g., its "coverage") is a topic worth investigation in its own right. We devoted some attention to it but restricted our investigation to aspects which bear fairly clearly on the evaluation problem.

2. Limitations and Restrictions

Since the area of information retrieval subsumes a whole host of different techniques, approaches, systems, and applications, it is important to emphasize the limitations of scope we have applied to our work. We have confined our attention to retrieval techniques applicable to very large collections of topically heterogeneous natural language messages -- of possible size from several hundred thousand to over a million messages. Such techniques are of the conventional "coordinate" document retrieval and the newer "associative retrieval" kinds. The boundary conditions we consider include the case when there is direct interaction between machine and ultimate consumers of information and where search times may be in the order of minutes or less.

We have not been concerned with techniques which try to provide detailed "factual" answers to highly complex questions but rather with techniques which retrieve stored message items. We have not considered a host of techniques involving extensive manual analyses of data or a priori coding -- syntactic analyses, semantic analyses, or special restricted query languages -- because these are not economically applicable to very large heterogeneous collections within the present state of the art.

The message retrieval application we have considered corresponds to a major portion of what is commonly referred to as "document retrieval"; namely, that portion which leads a searcher to a set of informative abstracts or precis. Our concern stops at that point, and in our evaluations we exclude the steps (which a searcher might well employ in practice) of proceeding from such condensed message representations to identification of relevant full-length documents.

In electing our methods of attack in this problem area, we have recognized that there are several possible reasons for "evaluating" retrieval performance, and that these reasons can lead to quite different evaluation methodologies. The possible reasons include: (1) to help identify promising directions for research on retrieval methods and techniques, (2) to help design better systems of a given kind which are expected to operate in a known kind of environment, (3) to help in the practical choice of system alternatives within a highly specific context, and (4) to help identify how to improve the operation of a given specific system in a given specific environment.

The main aim of the work reported in this report is (2), although we have also been concerned with (1) and with a capability of passing to (3) in our next phase of work.

We have focuses on evaluation measures which relate to user satisfaction. This being a laboratory evaluation, we have not been concerned with measuring important operational parameters such as computer and operating personnel costs, data acquisition and updating policies, levels of service, etc. Likewise, because the technology of computers is rapidly changing, we have not been concerned with details of efficient computer utilization or with search time minimization. We know, by virtue of having done it, that the kinds of processing required are feasible and reasonably economical using today's outgoing generation of machines (such as the IBM 7094-1401). Our present computer programs allow us to conduct twenty or thirty associative searches of our 10,289 message collections in about an hour using a relatively small computer (IBM 1401). Our assumption is that processing will be considerably more economical and convenient using the incoming generation machines (such as the IBM 360-50, GE 645, UNIVAC 1108, etc.) and that, therefore, the most important measurements we can make at this time are those having to do with user satisfaction rather than operating efficiencies.

In our experimental work, we have of necessity had to restrict our universe of consideration to particular indexings of particular message collections. Likewise, we have had to restrict our choice of search strategies to be tested and to make decisions of how best to allocate the effort of the human evaluators. In making these decisions, we have recognized a necessity to work toward two somewhat distinct over-all goals, one being an immediate desire for additional insight, the other being a long-term desire for systematic proof of results.

While the present report is centered on the key "evaluation" question, the discussion also goes deeply into some of the research areas inferred above. Our motivation was to gain maximum insight from observing the performance of the system and subjecting that performance to analysis and criticism. As a consequence, our experimental investigations have been in the nature of "broad spectrum" studies -- tests of many different aspects of the over-all system. Each result contributes

to increased understanding of the whole situation, but no single test is as definitive as it might need to be to provide final proof of a point.

B. Method

Our investigation of the laboratory evaluation problem has largely proceeded in a step-by-step sequence which is reflected in the section organization of the following material.

- * A necessary first step was to delineate and comprehend the problem under study, to estimate its dimensions, and to isolate the main difficulties. In part, this consisted of a review of the state of the art of user-oriented approaches to retrieval system evaluation, extensive first-hand communication with those most active in this area, a critical review of our own direct experience, and exploration of the new considerations required for assessment of the interactive situation.
- * Various studies of statistical procedures, mathematical formalisms, and data reduction techniques applicable to the evaluation problem followed.
- * The preparations for experimentation proceeded concurrently with the above studies and involved the machine indexing and associated computer preprocessing of the 10,289 message data base. Pertinent statistical parameters of language usage in this collection were gathered and analyzed. Machine and manual indexing vocabularies were compared in detail. Simultaneously, efficient computer programs were prepared for searching the collection and retrieving lists of associated words and/or prints of retrieved messages. Word association profiles for the 1,000 index terms in both the machine-indexed and the manually-indexed collections were prepared in both magnetic tape and printed book form.
- * Based on the results of processing search requests against the test data base, we studied the effects due to differing evaluation procedures, search strategies, methods of using associations, and use of different individual evaluators. Altogether, our data consists of the analyses of the results of over seventy separate search requests and over 14,000 relevance judgments (i.e., decisions as to the

degree of relevance of a given message to the concept expressed by a given request).

- * To provide the basis for interpreting and extending observations of this type to the more interesting "real-life" situation, we have done planning work, theoretical calculations, and small-scale experiments related to extending the association matrix methods into real-time, man-machine interactive retrieval situations involving the use of very large data bases.

Some of our findings and results have to do with message retrieval and others with evaluation strategy -- the former are treated in Subsections C and D below, the later in Subsections E and F. All of our results stratify naturally into various levels of specificity. Our most specific findings stem directly from test observations made on our experimental data bases; the most important of these are mentioned in the body of this report, while others can be found in the various Technical Notes referenced throughout. Our most general and interesting results are in the category of being "best considered judgments." They are predictive remarks which result from the whole of our experience. Subsections C and E below are devoted to such general remarks. Subsections D and F give brief summaries of our most cogent specific findings.

C. General Findings About Message Retrieval

1. Feasibility of Associative Interactive Retrieval Systems

"Second-generation"* retrieval systems with the following general characteristics are technically feasible and, within appropriate environmental settings, are likely to be attractive from an over-all performance viewpoint.

- * Large natural-language message base; 100,000 to over 1,000,000 short message items (abstracts, etc.) in ordinary English.
- * Fully automatic indexing; no classification or other intellectual analysis required of information entering the system.

* By "first-generation" systems we mean those based on punched- and manual-card technology and those systems created more or less by transfer of the punched-card methods onto computers. Large existing systems of this type impose two stages of interpretive manual analysis that are performed by persons other than the author or the requestor: indexing of messages and indexing of requests.

- * No special training or specialized procedures required of users -- communications of requests are in unconstrained English via keyboard or other simple display; no expert "system intermediaries" required; no need for consultation of "authority lists" or use of special logical formulas for request formulation.
- * Direct user confrontation with the information system. A typical request results in a two- or three-step user-system interaction where the user may refine or reformulate his request based on his selection from machine-computed and displayed word association for his request.

a. Technical Feasibility

The essential technical features of such a system, including automatic indexing, natural language query handling, large-scale use of statistical word associations, provisions for user-system interaction, have been built into our prototype 10,289 message system. This system was designed for experimental rather than production purposes and is tied to a generation of computers now becoming obsolete. However, we see no fundamental technical obstacle in the way of implementation of a much larger production-oriented system of this type employing real-time processing modes in place of the batch processing modes now used.

b. Performance Attractiveness

In our judgment, an associative interacting system has the potential to satisfy real needs of users in a rapid and direct manner, given an environment in which direct user access to machine terminals is possible. Specifically, for a very high percentage of the subject-heading type search requests processed (upwards of 90%), the user should be able either to home in on a high performance search request in two or three cycles of interaction with the system, or establish with a fair confidence that the desired kind of information is not in the file. We see no reason that the total process should take more than five or ten minutes, given a "reactive" typewriter, a time-sharing computer environment, or other convenient technology for user access to the system.

We submit that a direct user-system confrontation of this type is apt to be much more effective on the whole than conventional systems which rely heavily on services of intermediary personnel and which impose operational time delays (hours, days) between expressions of information need and response.

c. Economic Considerations

Feasibility is, of course, a separate matter from desirability in a specific context. As a general policy, we continue to recommend that a specific study of the requirements and economics of a given operational environment be conducted before any kind of major information system is implemented. We have not made a detailed study of the economics of an associative interactive system in comparison with that of a comparable existing system within a given environment. Higher costs for hardware and real-time system operation may or may not be offset by savings due to elimination of indexers and system intermediary personnel, and such costs must be weighed against benefits due to improved service levels in the context of the particular environment concerned. Our impression is that the cost of adding an associative message retrieval capability to a general-purpose "public utility" time-sharing system is relatively modest.

2. Role of Statistical Word Associations

Computed statistical word associations can be used in a variety of ways for improving the performance and flexibility of message retrieval systems. They are potentially useful in the searching process -- albeit in different ways -- both within the context of an existing (first-generation) retrieval system and within that of a second-generation system of the kind described above.

a. Present-Generation Systems

Within the context of an existing-generation retrieval system, statistical associations may be computed among the manually-assigned index terms. They can then either (a) be used to expand requests in an automatic fashion and thereby improve performance of a computer searching program, or (b) be printed in the form of a machine-compiled thesaurus of statistical associations, which is furnished to the system intermediaries as a supplementary aid to manual searching.

We have not studied the economics of such steps in great detail; however, since only additional machine processing is involved, we feel that the costs would be reasonable for many existing major systems given the large investments and operating costs already associated with these systems.

For detailed and specific requests, automatic associative request expansion appears to improve the density (i.e., fraction) of relevant items among those retrieved by a factor of two or more, assuming a setting in which there is no request renegotiation between system and user. The over-all result is improvement of both recall and precision of a given search, both by the factor of two. The factor represents the minimum improvement we have observed to result from simple automatic use of computed word associations to expand the vocabulary of a detailed request before searching for messages. Since our experimental association process is crude and approximate, we assume it represents a lower bound of the factor ultimately attainable.

b. Future-Generation Systems

The potential utility of association is perhaps even greater within the context of a second-generation associative interactive retrieval system. For example, the statistical associative methodology permits a machine to respond to a short subject-heading search request by reporting an estimate of the (single-step) retrieval performance expected of that request, together with a display of vocabulary terms highly associated with those in the request -- the words displayed are those present in the collection and most associated with the request in that context. The associated words become available for request reformulation, refinement, or generalization -- all under the direct control of the user. Finally, associative retrieval enables display of retrieved messages in sequence of decreasing value of a computed estimate of relevance.

There are two general situations in which the associations are potentially useful: (a) when the user is unsure how best to formulate his search requirements in the vocabulary used in the collection, and (b) when the user's search formulation is such that the measures of expected single-step search

performance are low. In either case, the user can proceed directly to reformulate his request using words from the association profile, without a necessity to retrieve or inspect actual messages. As mentioned, our experience indicates that for subject-heading queries at most two or three cycles of interaction are sufficient for convergence upon a high-performance search request -- if such a request can be formulated for the given collection.

The association facilities thus appear to be potentially quite useful within the context of an interactive time-sharing environment.

3. Role of User Mediation in the Search Process

Our exploratory investigations indicate that there are at least three capacities in which user intervention in the search process significantly improves performance of our system; intervention contributes positively:

- (a) In request reformulation for subject-heading requests, using words drawn from a machine-generated association profile, as indicated in (2) above.
- (b) In identifying concept-bearing word strings within full-text requests, by underlining or otherwise pointing to them.
- (c) In pruning machine-generated association lists derived from full-text requests, deleting unwanted words.

Each of these types of intervention appears to contribute to improving search performance.

A sharing of functions between user and machine appears to be appropriate -- for instance, we have found that while human pruning of association lists is valuable, it is better to retain the machine-assigned term weights in the pruned lists than to manually reweight terms. This is because of the superior capacity of the machine over that of the human to account for and take advantage of the statistics of vocabulary term occurrences and co-occurrences in messages in the specific collection at hand.

There are many residual uncertainties involved in foreseeing the behavior of real users interacting with an associative system functioning on line. However, the results of our analysis of the prototype have convinced us that pursuit of pilot operational evaluation of time-sharing retrieval systems is fully justified at this time.

4. Role of Machine Indexing

Our comparison studies of manual and machine indexing indicate that the quality of simple forms of machine indexing applied to short messages is considerably better than one's pessimistic expectations. The indications are strong that there are circumstances under which machine indexing can be both more economical and produce performance of equal or superior quality than certain conventional forms of human indexing. This should not be interpreted as a suggestion that machine indexing has been shown to be competitive with the high quality performance of a careful and expert indexer. We conclude only that the possibility of using automatic indexing in practical contexts is worth considering seriously when the obvious circumstances are met (machine-readable text economically available, investment in high-quality manual indexing difficult to justify, etc.).

D. Specific Findings Relating to Message Retrieval

We recapitulate here in outline form some of the main observations based on experience with our experimental data bases and discussed in more detail in the body of the report, especially in Sections V and VI.

1. Subject-Heading Requests

A short subject-heading type request can be classified by machine into one of two populations: Content-Bearing Units (CBUs), which are word strings which occur in the texts of the messages of our collection with certain statistical properties of "cohesion" and "repeatability"; and other strings (non-CBUs). We have found that:

- (a) The probability of a subject-heading request being a CBU is about 0.3.
- (b) A request which is a CBU has a high probability (over 0.8) of yielding a conventional search for messages with "satisfactory" performance, e.g., one which satisfies the minimum condition that the first three messages retrieved are all relevant, using modified coordinate search logic.
- (c) A request which is not a CBU may not be likely to yield a high-performance search, but will nonetheless result in a retrieved association profile which is virtually guaranteed to contain at least a few words with meanings pertinent to the concept expressed by the heading. Only very rarely (7% of the time) is the association profile

for a new request empty, forcing the user to start again from scratch.

- (d) Assuming an interactive process involving reformulation of requests using words drawn from the machine-produced association list, about 90% of all original subject-heading requests would result in "satisfactory" performance searches with at most two request reformulations being required.

The numbers mentioned above result from an argument which assumes NASA indexing statistics to represent query usages in our collection. However, changing the statistics would affect the rate of convergence but not the general result -- i.e., changing the numbers even quite considerably would result in perhaps one or three request reformulations being required instead of two as indicated.

2. Full-Text Queries

We have tested ten principal search options against full-text paragraph-long queries. In increasing order of over-all performance effectiveness, the four main ones are:

- (a) Modified Coordinate -- in which each message was machine-weighted according to the sum of the number of terms common to it and the original query. Messages are retrieved in decreasing order of their weights.
- (b) Frequency-Weighted Coordinate -- similar to (a) except that each message is weighted instead by the sum of the reciprocals of the collection occurrence frequencies of terms shared by query and message.
- (c) Reweighted Associative -- association lists for the query are printed out and inspected; terms are reassigned weights (positive, negative, or zero) by the user prior to actual message retrieval. Then, using all terms in the association list, retrieval is as in (a).
- (d) Selected Associations -- similar to (c) except that machine-assigned weights are retained. User confines his decisions to elimination of unwanted terms.

The associative options (c) and (d) outperform the nonassociative ones (a) and (b) by factors which vary from 1.5 to about 3, considering cumulative relevance points as a function of the number of messages retrieved varying over the range from 1 to 140 retrieved messages.

By the same criteria, (d) outperforms (a) by a factor of about 3 over the whole range. The total amount of relevant material among that retrieved is in about the same proportion for the first 140 messages. An exception is that option (b) outperforms (a) only in that it retrieves the same messages sooner, not in that more relevance points are retrieved. We have estimated an average total recall figure of about 50% for option (d), considering the first 140 messages retrieved.

3. Machine vs. Human Indexing

Our data enabled us to conduct detailed comparison studies of two specific indexing approaches: the original manual UNITERM indexing of our messages done by GE, and the automatic indexing we accomplished using a simple frequency-based criterion for selection of words from text. These studies indicated that:

- (a) Coverage of conceptual material in individual messages is about the same and is to roughly the same depth for both indexings (average of 3.5 concepts per message missed for manual vs. 3.1 for machine). Also, number of terms assigned per message is about the same. Machine-assigned terms are often identical with or close cognates of the manually-assigned ones.
- (b) The UNITERM indexing reveals numerous spurious term assignments -- i.e., index terms which describe concepts which are, in fact, not mentioned in the text of the corresponding message (average of 4.1 per message). Such terms might quite validly describe the contents of parent documents of which our messages are abstracts, but their presence creates a problem for the retriever who wishes to inspect abstracts as an intermediate step to document retrieval. The machine indexing does not produce spurious assignments of this kind, since all machine-assigned terms are contained in the corresponding abstract.
- (c) A much smaller-sized total vocabulary is required for the machine indexing to achieve the comparable coverages (999 terms vs. 4,824 UNITERMS).

These observations are clearly idiosyncratic to the particular indexing methods involved, and alternate and improved strategies exist for both manual and machine indexing. However, in our judgment, they provide a basis for recognizing that circumstances exist in which automatic indexing can be both feasible and preferable to the manual alternative.

E. General Observations Relating to Retrieval Methodology

1. Retrieval Parameters and Sources of Variability

An immense number of parameters must be considered when undertaking the evaluation of a retrieval system, and evaluation itself is meaningful only with respect to well-specified values -- or ranges of values -- of each of them. These parameters fall into four classes having to do with: (1) the user population and its characteristics, (2) the message collection and its nature, (3) how the collection is indexed, and (4) how searching is conducted. Real retrieval systems are of many diverse kinds, and tend to correspond to quite distinct typical parameter combinations. Even given a particular retrieval system in a particular environment, there are likely to be large amounts of indigenous variability for many of the parameters of that system. That is, it may serve users with quite diverse backgrounds; it may contain messages of quite different kinds and substance; it may be indexed by several different individuals; and it may be searched by several alternative strategies. If the results are to serve any useful purpose, evaluation experiments must be conducted and the results interpreted within a framework which clearly delineates which of these parameters are assumed to be given and constant, which are assumed to be variable for testing.

2. Absence of a General Evaluation Methodology

Because of the many kinds of variability, both among and within real retrieval systems, a general methodology for obtaining anything but a partial measure for comparative evaluation of retrieval system performance does not exist. That is, we are unaware of any methodology which measures more than one or two selected features of performance, and we consider it premature to try to create one at this time. The general approach for generating data relating to user satisfaction which seems most satisfactory to us for purposes of laboratory evaluation is based on the making of "relevance" judgments. However, it must be recognized that there are many features of customer satisfaction which do not have to do with relevance -- convenience of access to the computer, for example.

Even if analysis is based on relevance judgments, our studies of statistical procedures, mathematical models, and data reduction methods indicated that a wide diversity of defensible, rational, and interesting data analysis techniques had attractive features. Some of the candidate building blocks upon which a systematic evaluation methodology could be built included rank correlation methods, retrieval sampling procedures, multi-evaluator unbiasing methods, the use of various normalized precision and recall measures, etc. But it also became clear that each formalism built in certain assumptions about the nature of the data or the "objectives" of the search. As a rule,

assumptions about the expectations of users or about their expected reactions in stated situations are not available for testing in a laboratory context. Building such assumptions into the evaluation methodology or its formalism could obscure the truth and warp whatever insight might otherwise be gained. We concluded that empirical tests of an exploratory kind were most appropriate to the laboratory evaluation task at hand, and that they were best conducted outside any single formal framework. However, some form of discipline is necessary. We found it most helpful to summarize our results for interpretation in a form which forces (allows) the interpreter to make his own assumptions about the searcher's expectations and reactions. The reader will, as a consequence, find that the text of this report continually suggests alternate avenues of human behavior that might be encountered.

The reactions of users to the output of a retrieval process and the generation of judgments of what is useful or relevant are extremely complex behavioral phenomena. Techniques for measuring these phenomena are subject to the same kinds of uncertainty, variability, and ambiguity encountered in measuring other complex forms of intellectually guided human behavior. There are very few measurement methodologies which have gained universal acceptance in applied psychology except perhaps some applicable only to highly restricted subproblems -- and there is every reason to expect that this will continue to be true for the over-all problem of evaluating retrieval systems.

3. Elements for Message Retrieval Evaluation

We have assumed certain ranges of user, collection, and indexing parameters to be typical for the message retrieval application and have worked on a general approach to handling the remaining kinds of variability; namely, variability in search strategy (e.g., fully automatic or human-aided), variability in kinds of request (e.g., for specific content of request), and variability among users. To reduce the variability we have used both formal and informal strategies for:

(a) selecting requests, (b) sampling the collection, (c) eliciting relevance judgments from evaluators, and (d) combining, displaying, and comparing the results of evaluation.

4. Proof-Oriented vs. Insight-Oriented Tests

A large number of independent human relevance judgments is required to achieve any meaningful results, and economic considerations require that choices be made as to how available evaluator manpower is to be allocated. Proof-oriented tests involve concentration of evaluator effort into large-scale statistically valid investigations of one type of variability, with other conditions being held relatively fixed. Insight-oriented tests imply spreading of the same manpower over a number of investigative forays designed not to give conclusive proof

of a single hypothesis but insight as to how experimental variables interact with one another. While proof-oriented experiments are clearly desirable in specific contexts of operational evaluation, our experience indicates that the uncertainties inherent in the present state-of-the-art are such that there are numerous blind alleys, and laboratory evaluation must be flexible and consist mostly of insight-oriented tests.

F. Specific Observations Relating to Evaluation Methodology

1. Measures of Retrieval Performance

a. Precision and Recall

For laboratory evaluation purposes, we found these measures to be useful only under certain specialized circumstances; namely, (i) when there is a clear-cut cleavage between what is retrieved and what is not retrieved and when ranking among retrieved items is not important, and (ii) when an all-or-none criterion of relevance judgment is appropriate. While we have used precision and recall for certain evaluation tasks which do involve such circumstances, we have found that many other important tasks -- particularly ones connected with evaluating an associative searching system -- require more sensitive measures. We have found the precision and recall measures to be potentially useful for communicating crude summary data about system performance, but there is a concomitant danger of miscommunication because of failure to specify in detail the conditions under which the given figures are valid.

b. Other Summary Measures

Several other summary measures can be used to characterize the performance of a retrieval system; these include "normalized" precision and recall measures, Kendall's τ , Spearman's ρ , and the "M-V" rank correlation statistics. These measures, like precision and recall, boil down all performance data into one or two numbers. They do not readily allow for differences in depth-of-search requirement among different users and tend to be too gross for many evaluation applications.

c. Performance Characteristics Curves

We have found that as long as a retrieval system is likely to be used on different occasions with different search objectives in mind, the system's performance is more meaningfully characterized by curves or families of curves than by numerical measures which assume or otherwise incorporate a standard objective. The performance characteristics curves we have found to be the most useful exhibit plots of cumulative value of retrieved messages as a function of their rank of retrieval. The value of a retrieved message is that assigned by one or more evaluators. A single chart can exhibit curves for a single query or for average performance for an ensemble of queries, and can show performance for several different search options in juxtaposition. Inspecting such a curve, one can see how alternative search options perform with respect to variable depth of search requirements. One can see at a glance which options are best for those who want a few relevant messages, those who wish to compile a bibliography, or those who want mostly to browse.

d. Advanced Performance Measures

It is possible to identify special-purpose performance measures which are appropriate for use if search objectives of a system can be specifically identified. We have developed three such measures: a "Sliding Ratio" statistic which embodies a free parameter which can be set according to depth-of-search requirement, a "Browser's Statistic" appropriate for evaluation when what is desired is maximization of probability of a single success in sequential search, and a "Cost Matrix" method of measurement which allows great flexibility in identifying search objective criteria. We have found no need to use such measures in our laboratory evaluation work, however, for the Performance Characteristics curves have been completely adequate for our purposes.

2. Judging Relevance

a. Levels of Relevance

Our experience with the message collection studied indicates that in some circumstances only two levels of relevance need to be recognized in making evaluation judgments, but that in others multiple levels of relevance

are valuable. In particular, we have found that for paragraph-long requests there are exceedingly few messages with very high relevance, but a substantial number with varying degrees of partial relevance. Under these conditions, making relevance judgments on a graded scale (say 0 to 4) is natural and results in less arbitrary decisions than insisting on all-or-none judgments.

b. Combining Results of Evaluators

We have developed various mathematical methods for combining and unbiasing the results of several evaluators who simultaneously inspect the output of a given retrieval run. However, we found that redundant use of evaluators was both unnecessary and uneconomical for our purposes of laboratory investigation. Limited experiments using single and multiple evaluators indicated that significant relative differences between search options could be identified adequately using a single evaluator, and that evaluator effort was best expended by investigating additional query-message pairs rather than by replicating judgments. We are unsure, however, of whether or not use of multiple evaluators will be desirable under situations of operational evaluation, where differences in performance among search options may be smaller than those we have been dealing with so far. Fortunately, we found it easy to test whether a single evaluator could act as the agent or representative of a panel; comparatively modest experiments sufficed to establish the validity of this economy of effort. Thus, the prospect of generating performance characteristic curves over a board base of queries was not necessarily forbidding.

SECTION II

THE MESSAGE RETRIEVAL EVALUATION PROBLEM

The purpose of this section is to provide a context for the discussion of our work which occupies the remainder of this report. Subsection A is devoted to specifying the kind of retrieval systems we are concerned with evaluating, in terms of user parameters, collection parameters, indexing parameters, and search parameters. Subsection B is concerned with approaches to evaluation. We try to make clear our viewpoint toward laboratory evaluation, and our motives behind this viewpoint. Part of this discussion is concerned with the requirements generated by new-generation "interactive" retrieval systems for quite novel approaches to the evaluation problem. Subsection C is concerned with the key construct of "relevance" which underlies our approach to evaluation, and brief reference to work of others is made in Subsection D.

A. Identification of Retrieval Parameters

Information retrieval, message retrieval, or document retrieval are labels that have been applied to a large variety of quite different activities conducted with quite different objectives. These range from manual search for an unknown reference in a library to machine search for a highly specific data item in a military query system. Before one can reasonably discuss the evaluation of a prototype system, it is necessary to specify the bounds on four main types of parameters in order to delimit the area and scope of inquiry. These parameter classes are:

- (1) User Parameters: parameters which describe the population of users of the system, including their background skills and kinds of interests, the needs for information they expect the system to satisfy, the kinds of requests they expect to pose, and the kinds of responses they would like to have.
- (2) Collection Parameters: parameters which describe the kinds of informational items in the data collection, whether numerical data, messages, abstracts, or whole documents, the number and lengths of these items, the fields of interest they treat, and the methods by which they can be used to obtain answers to questions.
- (3) Indexing Parameters: parameters which describe the mechanisms and conventions by means of which the informational items are indexed, descriptions of procedures for assigning codes, artificial descriptors, classification schemes, or elements drawn from natural language, etc. Characterization of relevant parameters

of the index term set, including number of terms per message item, frequency distribution of terms, etc.

- (4) Search Parameters: parameters which characterize the conditions and methods by means of which users can have access to the informational items, including machine or man-machine search options available, access time consideration, whether capabilities for automatic term association are available, etc.

We have focused on a restricted region within each of these parameter areas, designed to correspond, on the whole, to the military "automatic message retrieval" problem which we have discussed in previous reports and Technical Notes.* The viewpoints we have taken with respect to each of the parameter areas are as follows.

1. User Parameters

We assume that the user population consists of individuals with at least a fair over-all level of technical training in the subject areas dealt with by the collection, but not training which is necessarily specific to the information collection at hand. We assume that there is a minimum number of such potential users -- say, a few hundred -- and that they are primarily engaged in some area of activity (other than maintenance of the retrieval system) from which they derive the need to retrieve information from the common store which they will personally analyze and put to use. For example, they might be a group of research workers, or perhaps they are experts in various areas of technical intelligence. These users are assumed to have a variety of kinds of needs for information, ranging, on the one hand, from a desire to locate a well-defined and highly technical unit of data (maximum thrust developed by a TX-354 Rocket Engine) to a request for an exhaustive bibliography covering some broad area (a bibliography of references having to do with design and performance of rocket engines). We assume that an important need these users have is to conduct searches for detailed information which is likely to be scarce in the collection and packaged into messages in unpredictable ways. Such a need is likely to be felt by intelligence analysts, for example, and is particularly difficult to satisfy using conventional retrieval methods. We assume that the user's time is limited and valuable, so that, in general, he will want to look at anywhere from a few message items to a few hundred. We assume that completeness of response is often of importance to him, but not at any cost of his effort. We assume that he is willing and able to interact directly with a machine to get his answers if this is easy and does not require specialized training, extra work, or inconvenience on his part.

* See Arthur D. Little reports (1) and (2).

2. Collection Parameters

We assume that the collection consists of a large number (up to more than a million) of brief, independent, natural-language message items, each of which might be from 20 to 100 words in length. They could consist of intelligence or command and control messages dealing with a technical area. The message items might carry references to longer documents or reports, and in our present 10,289-message experimental collection they consist of abstracts of documents dealing with aerospace technology. Although message items may contain graphics, illustrations, or equations, it is assumed that the bulk of the information is carried in the English text. The technical area dealt with by the collection is assumed to be sufficiently narrow that a fairly large number of highly technical terms will come to be used in talking about it; the area is also assumed to be sufficiently broad that the language used in the messages is not the private code system of a closed community but rather is best left to be unconstrained natural English -- although technical or military style might exercise major influence.

In general, we assume that a typical user inspecting a message item will either be able directly to identify whether it contains an answer to his question, or whether the parent document it references (if it is an abstract) has a high probability of containing the answer. We do not consider the aspects of search connected with reading parent documents, but are concerned only with evaluating the portions of the process which terminates with the identification of presumably relevant short "message items."

3. Indexing Parameters

We assume that, for reasons of economy, speed, efficiency, etc., indexing is to be done by machine through direct automatic processing of the texts of messages, and that there is no human mediation in this process. In other words, indexing is to be accomplished via the use of natural-language descriptive words and phrases encountered in text and selected through appropriate machine-implementable criteria. The indexing vocabulary and procedures are assumed to be designed to provide -- within economic limits, of course -- detailed characteristics of the contents of the messages so as to provide maximum possible selectivity in searching. In terms of existing systems, this means that a typical message will be indexed by twelve to thirty descriptive tags.

4. Search Parameters

We have focused on a narrowly defined class of searching options for the message retrieval application, all of which are machine-based, and some of which are dependent on limited forms of human mediation in the search process. The main strategies we have been concerned with in

our experimental work have already been identified in Section I and are described in detail in Sections V and VI.

B. Approaches to Evaluation

1. Main Considerations

Evaluation implies the weighing of alternatives in the light of a desire to meet certain objectives. We have focused on the following objectives of evaluation:

- (i) to help design and test better kinds of message retrieval systems of the general types we have described above;
- (ii) to help identify promising directions for further research on associative message retrieval methods and natural-language processing;
- (iii) to help provide guidelines for the choice of system alternatives which would be applicable in a specific operational environment.

Most of our emphasis has been on (i), some on (ii), and comparatively little on (iii). The alternatives we are interested in weighing are primarily ones of search strategy. Much of our work has been addressed, then, to comparing search strategies given our assignment of emphasis to objectives (i), (ii), and (iii).

Evaluation of a retrieval system can be based on consideration of two main categories of costs and benefits: (a) those which accrue to operators of the system, and (b) those which accrue to its users.

Operating costs include costs of various categories of operating and administrative personnel, machine purchase or rental costs, data acquisition, and preprocessing costs, data transmission costs, and general overhead. Use of one retrieval search option as compared with another can affect these costs in several ways such as (a) requiring more or less computer time for searching, (b) requiring a larger or a smaller machine, (c) requiring more or less manual data preprocessing effort, (d) requiring more or less time for the manual part of search generation and renegotiation and interpretation of results, and (e) requiring use of higher or lower calibre personnel for these functions, etc.

Costs accruing to the user include (a) money paid for a search, (b) costs connected with delay or inconvenience in access to the

facility or in getting the results, and (c) time and effort required for formulating an inquiry, communicating it to the system, possible renegotiation of the inquiry, and perusal and analysis of results. Finally, (d) the benefits to the user are associated with the successful discovery of useful relevant material and are somewhat less tangible and more difficult to measure. Use of different search options can clearly have major impacts on each of (a) - (d).

We have concentrated most of our evaluation effort on costs and benefits accruing to the user, particularly on items (c) and (d). We feel that this is an appropriate choice given that our main objective is design of improved kinds of message retrieval systems, not optimizing the operation of a particular installation. With regard to operating costs, we assume that the computer requirements and total cost per search of the search options being tested are roughly comparable.

As to user costs, in the absence of a specific operational environment it is not realistic to conduct detailed comparisons of cost paid per search or costs connected with access. We do assume that the dollar cost per search is reasonable and justified with respect to the objectives of the users, and reasonably independent of the search strategy employed. As to access convenience, we believe that it is realistic to orient ourselves to the presently coming generation of time-shared bulk memory computers. That is, we assume that users have direct access to remote console terminals, and that normal response time between posing an inquiry and getting a list of associated words is something under thirty seconds, and that it takes less than three to five minutes for the machine to translate an edited association list into a typewriter printout or cathode-ray tube display of presumably relevant messages.

We regard user time and effort to be of great importance in the present context and, for this reason, assume that queries to the system are either in natural English or very lightly edited form -- we do not consider search options where the user consults "authority lists," employs complex Boolean search prescriptions, etc. Negotiation between user and machine, if any, is assumed to consist of selection of terms from a machine-generated thesaurus listing for the given query at hand, and possible reformulation and resubmittal of a query. Some user effort is assumed to go into reading association lists and reformulating queries. However, the bulk of it is assumed to be concentrated in the readings of the retrieved message items, and in assessing their relevance to the query at hand.

2. New Directions in Retrieval System Evaluation Methodologies -- A General Discussion

By an automatic retrieval system we mean an integrated configuration of equipment and procedures, both manual and machine, for

retrieval access to a data base consisting of natural-language messages. Generally, such a configuration is made available to help a well-defined group of people accomplish identifiable storage and retrieval functions defined by the needs of a real operational environment. In creating the next generation of such systems, powerful techniques can be brought to bear: the full capabilities of very large computers, the human engineering advantages of direct-access consoles, the "conversational mode" of man-machine interaction, and the advanced methodology of mathematical systems modeling. The resulting creations are far from being simple ad hoc devices for performing crude functions; rather, the tendency is toward complicated and venturesome conceptions for performing ambitious tasks. The evaluation or appraisal of the performance of systems of this type, be it in a pilot-operational context or while the system is in its prototype stages, poses new and challenging problems.

The design of a fair test for these "second-generation" retrieval systems requires several departures from traditional evaluation methodologies. Because the system designs embody the notion of user interaction in searching and because they embody highly mathematical processing methodologies and are provided with ample speed, memory, and computing power to carry them out, it is necessary to regard these systems as capable of doing, as a matter of course, things which were regarded a few years ago as wholly impractical. Therefore, there exist several matters of urgent interest in the practical context of a real evaluation which did not have to be considered in the past.

The capability of automatic message retrieval systems to provide ranked output is a case in point. Without the availability of an adequate mathematical formalism or data-processing facilities of sufficient power, the idea of responding to a query by displaying first the most "relevant" document, then the next more "relevant" document, and so forth, is theoretically interesting but practically impossible. However, the theory and technology of retrieval searching has advanced considerably since the days of "peek-a-boo" coordination, and the provision of ranked output is today a standard feature of several operational retrieval systems.

As a consequence of this development alone, it was clear to us that there was a need to reappraise some of the prevalent notions as to how a retrieval system is to be evaluated. Evaluation methodologies that presume the system will select a set of messages (rejecting the rest) are based on an assumption that simply may not apply to the newer automatic systems. If the messages are regarded as ranked, then the system does not actually "retrieve" any clear-cut set of documents nor does it reject "irrelevant" documents -- rather, the whole concept of a dichotomy between retrieved and nonretrieved messages disappears until it is reintroduced by de facto human cutoff action during the course of practical search operations. Since the system does not select and reject, it should not be judged within a framework that assumes it does or ought to.

We notice at once that the new technology changes another ground rule of evaluation; in many systems, the number of documents "retrieved" is a significant parameter of the search event. Once a two-term coordinate search has been submitted, for example, the number of documents retrieved is entirely determined. "If it is zero, the search was wasted; if it is 5,000, the number is too large" are typical statements of an evaluative kind that could be based on this parameter. But in principle when the output is merely ranked, the "number" of documents retrieved is not a specified feature of the search event. It is certainly not a property of the retrieval system. Rather, it is an expression of choice and option on the part of the requestor who can easily be thought of as reading as few or as many of the ranked messages as he cares to. Naturally, he will not care to read the entire collection, so there is a premium on placing messages that are probably relevant at the head of the output list or near it. But there is also the compensating factor that people who are, in fact, interested in exhaustivity of search will also be motivated to read further down the ranked output list and thus, in fact, "retrieve" more documents than those not motivated. In short, new concepts of what the user is doing and of what the system is doing have necessitated the development of new views of what evaluation should consist.

Another closely related case in point is that the new-generation systems are designed to be "conversational," i.e., they are designed for iterative rather than one-shot searching. However, almost all existing conventional measures of system performance are designed to evaluate the performance of a single-shot searching process: a query is put to the system, and the response consists of retrieved messages or documents. This is an inappropriately narrow viewpoint to take of an interactive system. The crucial question to be asked in evaluating such a system is not how good the average performance of a one-shot search is; rather, it is how quickly, how easily, and with what certainty can a user home in on a high-performance search.

Another major influence requiring extension of the power of applicable evaluation methodologies has been the lavish multiplication of feasible search alternatives that occurs when an interactive facility is provided. When one actually has such a system in prototype test operation, it is commonplace and natural to try out several optional strategies for every search conducted -- in the interests of gathering evidence that suggests which options are preferable under what conditions. It is an evident consequence of the great ease with which one can change the searching protocols that one is dealing not with one system but with many. So far as future automatic message retrieval systems are concerned, the relatively comfortable days when one quite rigid system was measured against one rigid standard have thus passed.

The proliferation of alternatives is most severe, of course, in a research context -- where options can be explored without concern for immediate practical implementability -- and it would be tempting

to suppose that the evaluation problem would be far simpler in a practical operational situation. But this is probably wishful thinking: given high-power machines, it is simply too easy to provide a full panorama of alternative search and interaction options for exploitation. The incremental cost of providing and of using a variant option is small in comparison to the total investment in the system. In thinking about evaluating operational systems of this kind, it is a foregone conclusion that there will be a fairly large number of strategies built into the system and available to users for attaining all kinds of alternative search objectives under different conditions. Thus the nightmare of evaluating a set of systems is here to stay, not only in the "retrieval system research laboratory" but in practice as well.

All the complexities of assessing different search objectives, of evaluating the requestor's capability to choose a good search strategy (together with the system's capability to guide him in such choice), and of determining in balance whether all the expense, labor, and effort is really helpful and worthwhile, join together to comprise the problem of "evaluating an automatic message retrieval system." We consider the work reported hereunder to be but an initial attack on this problem.

C. Concerning Relevance

Our approach to performance analysis is to estimate, usually comparatively, how well a retrieval system is likely to satisfy a user in terms of the effort it takes to find relevant messages among those retrieved. To do this we make certain measurements and then interpret the data generated by these measurements in various ways. These interpretations depend heavily on the construct of relevance; that is, relevance of a retrieved message to the intent of a query.

The construct of relevance plays a crucial role throughout the discussions in this report and in almost all other published accounts dealing with the methodology or practice of retrieval system evaluation. Unfortunately, however, the construct has too frequently been either left undefined or treated in a confusing manner. In particular, a commonly recurring source of confusion has been the ambiguous use of "relevance" as a name for both an intellectual construct and for a particular measure applicable to that construct (the measure which we later call "precision").*

* Failure to adhere to a reasonable degree of operationalism in discussions about relevance has led to more than one imbroglio in print. See, for example, Taube in ref. (3). The concept is discussed from various points of view in refs. (4) and (5).

1. Public vs. Private Relevance

We feel that it is important to begin by drawing an important distinction between two related but distant intellectual constructs, which we will call private relevance and public relevance. By private relevance, we refer to a class of very specific events which take place at particular spots in time and space. Namely, suppose we are given a particular individual J with a particular need for information N at a given time and space X, and we know that he is confronted at that moment with a particular item of information M. By private relevance we mean the degree D to which M satisfied J's need N at time and place X. Determination of this degree is the event of relevance. An event of private relevance is thus highly subjective, ephemeral, and unique, although it is well-defined and meaningful for the individual concerned at that moment. Such events are to a large extent incomparable. Not only may the needs of different individuals be of necessity different, but a given need of a given individual is bound to be different at different times, and for that matter the individual himself is likely to be undergoing change.

It should be noted that the construct of private relevance does not require that the structure of the need N be observable. An individual can merely indicate that a particular information time M either does or does not satisfy his need N without saying anything about how or why, and without in fact otherwise indicating what his need N is.

Ultimately, the data which we and others use as a basis for evaluating retrieval performance consists of the outcomes of multiple events of private relevance. Public relevance is a construct which refers not to a particular unit event, but to over-all statistical properties of certain classes of events of private relevance. It is a grosser construct, but much more useful for the basic task of evaluating retrieval systems; that is, relating how satisfying the output of a system is with respect to specification of information requirements which are encoded as input to that system. Specifically, suppose we are given an ensemble of individuals (say, some population of the users of the retrieval system) E, an information need as expressed by a written statement N', a representative set of time and place circumstances X', and a specific information item M. Any given determination by a specific individual belonging to E of the degree D to which M satisfied his own understanding of N' at place and time X is a finding of private relevance. By public relevance, we mean a statistical property such as the most usual or expected reaction by members of E to M, provided that they are given N' under general conditions X'.

This report is primarily concerned with public rather than private relevance, and with its measurement and estimation. That is, needs for information are assumed to be expressed in a public format (i.e., queries are put down in writing), and we are concerned with the statistics of multiple events of private relevance, not with the

microstructure of one such event. Two unavoidable limitations are therefore associated with use of public relevance, and should be borne in mind. First of all, an individual may be incapable of doing much more than approximately characterizing his need N in a verbal statement N'. Even if he thinks about it, certain aspects of his real need are apt to be reflected poorly or not at all in his written expression of that need. Secondly, judgments or findings based on statistics of public relevance are bound to be wrong for a certain proportion of specific events. That is, what is relevant for the many may, in fact, be irrelevant for the (very important) individual specifically at hand and conversely.

2. Operationalism and Performance Measures

As just indicated, we have a general notion of what we mean by public relevance. To give the construct of public relevance a more precise meaning, it is necessary to state the exact measuring and statistical procedures to be used in a given circumstance -- i.e., to specify an experimental design. One of our objectives has been to construct and analyze various measures for retrieval system evaluation, such that these measures reflect the degree to which a retrieval system performs in terms of interpretations which are formulated using the construct of public relevance. Section IV is devoted to a discussion of such measures. We have recognized here a need for operationalism. We adopt a procedure for measurement and evaluation, and then discuss the behavior of various retrieval systems with respect to such measures and procedures. Such procedures and methods, in fact, mean no more than what they measure, and all discussion in terms of "relevance" and "user satisfaction," etc., are interpretations based on our notion of what we intend to measure, or what our measures are intended to estimate, nothing more or less.

D. Previous Work

Annotated bibliographies and critical discussions of previous work on retrieval system evaluation procedures are available, and we will not attempt to review this past work here.* Many important aspects of evaluation in working environments are discussed from numerous points of view. We have found much of this literature to be valuable in pointing to many of the aspects, facets, and pitfalls of evaluation, and

* A rather comprehensive annotated bibliography has been prepared by Madeline M. Henderson of the National Bureau of Standards (6). See also items (7)-(12). An excellent source of references is the "Literature Notes" section of American Documentation, which may contain a dozen or more abstracts or reviews of articles on the retrieval evaluation problem in a typical issue.

well worth reading for this reason. Nonetheless, most of the previous work applies to boundary conditions other than those we have assumed for the message retrieval application, such as to compare modes of human classification and indexing, etc. As discussed earlier, we have found that most of the evaluation measuring techniques discussed in the literature are inadequate for the main application at hand -- laboratory evaluation of the performance of an experimental prototype of a second-generation associative searching system.

SECTION III EXPERIMENTAL DESIGN CONSIDERATIONS

In this section we comment on some important selected considerations having to do with the design of retrieval system evaluation experiments. These considerations include collection size and sampling strategies, the economics of evaluation, selecting requests, making relevance judgments, combining the judgments of different evaluators, costs of evaluation experiments, and the distinction between insight-oriented and proof-oriented experiments. Some of our remarks are general; but most relate to how we have dealt with these considerations in our own work.

This section deals mainly with various steps in an evaluation process leading up to but not through the analyses of data obtained from experiments. A discussion of various procedures for data analysis and treatment of performance evaluation measures is reserved for Section IV.

A. Collection Size and Sampling Strategies

There is a serious problem of scale connected with the laboratory testing of retrieval methods which are intended for use with very large collections. Search options and processing methods which work well for a collection of 500 messages may be uneconomical or, for all practical purposes, impossible with a collection of 500,000 messages; the same is true for many evaluation techniques. There is no guarantee whatsoever of extendability; the options which work best when applied to a small collection might be far from best when applied to a much larger one.

It appeared to us that there were three general ranges of collection size within which we could conduct experiments: large (above 100,000 messages), medium (10,000 messages and above), and small (say, under 500 messages). We and others have already reported on initial exploratory evaluation work performed on a number of collections in the small range.* We rejected the notion of working further with small collections, particularly because we felt we could not comfortably extrapolate the results of evaluation -- even though small collections are highly attractive from the viewpoints of processing economics, manipulative ease, and the potential for thoroughness in the evaluation performed. In Section IV, we discuss the possibility of designing mathematical models to extrapolate evaluative results from smaller to larger collections, although so far we have attempted to model only simple nonassociative search strategies in this manner.

* See (1).

We rejected the large range in the laboratory phase of work reported here because of the awkwardness, lack of flexibility, and high cost experienced in handling such a large body of data. We compromised with our 10,289-message collection, feeling that we had thus purchased some, but not too much, of both evils. The collection is smaller and its processing was more expensive and cumbersome than we would have liked; on the other hand, this collection is more than twenty times larger than collections used previously for systematic evaluation of associative search methods,* and is still more than sufficiently flexible for our purposes.

Perhaps the most serious problem encountered when using any collection other than a small one is that in practice the evaluators of the retrieval output cannot know exactly which messages in the collection are relevant to a given query. That is, given a query, it is not practically possible for human evaluators to read all 10,000 or 500,000 messages in the collection to find out how many are, in fact, relevant but not listed with high value in the computer's output.

This practical roadblock forces the adoption of some kind of sampling strategy, so that a human evaluator is only required to read and judge the relevance of messages in a sample which is much smaller than the collection as a whole. The objective, naturally, is to create samples for evaluation in such a way that an accurate picture can be determined of how well a retrieval system is doing in terms of ranking messages with respect to their relevance to a given query. However, the problem is well known to be more complex than it appears. For example, random sampling from the collection is unsatisfactory because, for a specific query, a random sample is apt to contain only messages with very low relevance. On the other hand, a sample of messages selected from among those with high machine-assigned relevance weights may not reveal what the machine missed. To overcome this particular problem, several experimental designs for doing "intensive" sampling have been proposed, modified, tested, and criticized.**

* To our knowledge, the largest collection used for evaluative associative retrieval experiments was 450 abstracts (13). Other experiments with very limited data were reported by Dale and Dale (14). H. E. Stiles has reported work on much larger collections, but not of a systematic evaluative nature. Curtice and Rosenberg have also reported associative experiments dealing with 800 documents (15) and have since increased to 1800 documents, but again the evaluation aspects have not been emphasized.

** Calvin Mooers first suggested sampling in (16) although the scheme suggested there has serious flaws. Modifications of this method were suggested by Fels (17) and Bornstein (18). See also Bryant (19).

We have found the following relatively simple procedures to be entirely adequate for the purposes of comparing various search options in the prototype system:

- (i) For a given query, we construct an "enriched" sample by forming the union of the topmost portions of the output lists obtained through using several retrieval search options (preferably ones which are as independent as possible in their retrieval logic).
- (ii) We use a random sampling technique to construct a "control" sample for the same query.

The evaluator(s) goes through both samples, making judgments of the degree of relevance to the query of each message encountered. We have found that an enriched sample obtained in this way has enough relevant messages in it to give a good indication of the performance characteristic (cumulative amount of relevant material as a function of rank) of a given retrieval option. The control sample, on the other hand, is used to give an estimate of the proportion of relevant messages which is actually included in the enriched sample.

In one series of experiments, for example, we used six search options to retrieve different message rankings for the same basic query, and the topmost 120 or so messages from each of the six lists of retrieved messages were merged to form the enriched sample. A typical enriched sample formed this way had less than 300 different messages, indicating that the search options were far from independent. We have found that for a short subject-heading request (2 or 3 words long) 80% or more of the relevant material in the collection is apt to be in an enriched sample obtained this way. For long, highly specific requests (a paragraph long), our enriched sample typically contains between 40% to 70% of the relevant material in the collection. More details on these results are given in Section VI.

B. Selecting Queries

By a query we mean a statement written in English which describes the information being sought. It is assumed to be the most complete public description of a given information need that is available. We find it desirable to distinguish between the query and a search formulation, the latter being the input to the machine portion of a retrieval system and possibly more condensed than the query.

In our present context -- one of experimental laboratory evaluation -- the experimental message collection we have assembled is too small and too old (the most recent items are dated 1962) to be of much

current interest to real users. We have, therefore, had to adopt strategies for selecting representative queries. The following paragraphs are concerned with some of the main considerations connected with the selection of queries; the specific tests we have done are described in Section VI.

1. Specificity and Topical Content

Two general characteristics of a query which have a major impact on retrieval performance are its topical content and its degree of specificity or generality.

A message retrieval system of the type we are concerned with may deal with literally tens or hundreds of thousands of distinct topics. Some of these will be covered very extensively -- i.e., will be the central theme of hundreds or even thousands of messages, but most will be touched on only peripherally -- i.e., mentioned in a subsidiary capacity in only one or two messages.

A short query, one or two words long, will be specific or general with respect to a collection depending on whether or not it deals with one of the central topics of that collection. In our collection, about 500 messages deal directly with "heat flow," about twenty with "plasma jet apparatus," and some six with "dislocation theory." However, a long detailed query, say, specified by a paragraph, will always be highly specific, whether or not it deals with topics central to that collection -- the problem is that it may be so specific as to exclude the entire collection.

It makes little sense to pose detailed queries dealing with one topic area to a collection which treats that topic area peripherally, if at all (such as posing a detailed question about dental surgery to a collection having to do with rocket technology).

Several definitions are possible of the degree of specificity of a query, and we have been concerned with modeling this characteristic in work reported elsewhere.* This model provides an operational definition of specificity in terms of the statistics of usage of query language expressions in the message collection being searched. Roughly speaking, a longer query is more specific than a shorter one. Thus, the query "heat flow" places many fewer constraints on what is required than "axial heat flow in turbine impeller shafts." For a given length, a query containing rare words and rare word strings (rare in the collection, that is) is more specific than one containing frequent words and strings. Specificity in this sense depends very much on the

* See Technical Note CACL-18 (2).

message collection at hand, and a query which is highly specific with respect to one collection may be very general with respect to another. Specificity is thus determined by a combination of length of query and relative uniqueness of its vocabulary and word string usages in the given collection. In our collections, "invidious intersusception" is much more specific than "heat flow" and so is "heart attack."

2. Kinds of Queries

a. Subject-Heading Queries

In most large-scale documentary information systems, a very important kind of query is the subject-heading request; i.e., the user expresses his need for information by means of a short descriptive string of words (Examples: "Pocket Re-entry Friction," "Thermal Control," "Cyclotron Radiation," or "Turbulence Studies"). Such requests are in effect the only ones which can be handled by the traditional classification-based library systems. In a typical classification-based system, many of the subject headings will be of a relatively general nature and have large numbers of items posted to each. Only very general queries can be accommodated. In a typical machine-oriented document retrieval system, on the other hand, indexing is typically done in much greater depth; there are many more subject terms and many more postings per document. It becomes possible to process very long and detailed queries. However, even for these systems there are at least two types of circumstances for which it will be desirable to handle subject-heading type queries:

Circumstance 1: The subject heading is the best available statement of what the user wants (or perhaps is the only available statement): it is neither too specific nor too general, and its inherent ambiguity mirrors ambiguity in the requestor's mind over what he wants. This circumstance might hold, for example, if the requestor is an analyst interested in compiling a bibliographic list of all references dealing with "Rotor Turbulence Studies."

Circumstance 2: The requestor does not start out by knowing how best to formulate his information need, and wishes to secure assistance from the information system in the process of query formulation. In this kind of circumstance, the user might well begin by posing a subject-heading request which is the abbreviated and perhaps approximate formulation which first comes to his mind. Later, he can modify or sharpen his request, on the basis of information fed back by the machine. For example, a requestor who begins

with the query "Rotor Turbulence Studies" might be interested only in highly specific information dealing with "Helicopter Rotor Vane Stability Analysis." However, he might not easily think about asking for it that way until he sees the words "helicopter," "vane," "stability," and "analysis" in a machine-furnished association profile.

b. Full-Text Queries

A second class of queries of importance for some applications consists of paragraph-long descriptions of what is wanted. We have noted previously that for certain retrieval applications, such as the correlation of contents of intelligence messages -- each new message may be processed through the system as a query before it is added to the data base.

In our experimental work, we have dealt with:

- (a) short subject-heading queries dealing with general topics; i.e., those central to the collection;
- (b) short subject-heading queries dealing with specific topics; i.e., those somewhat peripheral to the collection;
- (c) paragraph-long queries dealing with constellations of both general and specific topics.

We have omitted consideration of inquiries described in a unique way by addresses which are available in memory, such as author and date, message reference number etc. This case could, in fact, be subsumed under (b) if relevance criteria were so formulated that, in case of such a kind of inquiry, only the message which is being requested has positive relevance, all other messages automatically having relevance 0. We, nevertheless, have omitted such inquiries from our investigations because they can easily be handled from a practical point of view even though they are comparable on a theoretic plane.

3. How We Picked Test Questions

We wished to match our test queries to the topical areas treated in our experimental collection. As described in Section V, the general area of the collection is aerospace technology, with particular emphasis on engines and flight propulsion.

In the absence of additional guidelines, we wanted to use test queries representative to the area of technology of the collection, and of both the specific and general types mentioned previously. For this purpose, in several of our experiments we have used other operational retrieval systems -- ones dealing with the same body of technology over the same time span -- as sources for queries.

We have based a portion of our investigation of short queries upon the subject-heading vocabulary of the NASA document collection employed in the STAR (Scientific and Technical Aerospace Reports) periodical. For example, two members of one of our samples of short queries were "Control Valves" and "Ceramic Coatings." These are also headings used for subsections of STAR, and from month to month various publications are listed under them. Since these expressions were selected by human subject-matter specialists to function as retrieval keys, we reasoned that they too would be appropriate queries to pose to our experimental systems. By sampling and testing headings from STAR, then, we hoped to learn something about the over-all performance of our system for subject-heading queries. This line of reasoning is developed in detail in Section VI.

Some of our long detailed queries consisted of entire abstracts drawn from the Propulsion Systems section of the TAB (Technical Abstracts Bulletin) publication of DDC (Defense Documentation Center). We reasoned that an abstract provides a description of subject matter which is highly specific but at the same time realistic. Our view was "here is a real message which must be of interest because someone has written it; let us use the retrieval system to search in our collection for other messages pertinent to the topic specified by this one." We believe that this viewpoint about searching is likely to be quite common in certain environments; for example, when dealing with intelligence messages where correlation of new information with that previously received is of paramount importance.

C. Making and Recording Relevance Judgments

Another important set of experimental design considerations have to do with how relevance judgments are made and recorded. That is, how is an evaluator (or panel of evaluators) instructed to go about assessing the relevance of messages? How is he instructed to set down his assessments so that they can be conveniently analyzed? We have

investigated and/or tried several possible procedures at one time or another and briefly comment on these in this subsection. Discussion of a set of closely related considerations -- how to analyze the data resulting from relevance judgments -- is reserved for Section IV.

Regardless of which procedures are followed for recording judgments, in almost all of our work we have expected evaluators to base their judgments ultimately on their understanding of the public content of a written query. They have been asked to take the view that such a query is all that is known about an information need, even if it appears to be either overly specific or ambiguous to him. The evaluator is expected to assess relevance on a conceptual basis; not, for example, on a simple basis of matching word strings.

We refer to the results of the evaluators' processing of a sample of messages as the master list, since this defines an absolute standard against which the outputs of various retrieval options are to be compared. The sample may be the results of retrieval, random selection, etc.; in most of our work sample size has ranged between 150 and 300 messages. In the following paragraphs we first discuss various forms the master list can take, assuming that there is only one evaluator, and then pass on to the case when a panel of judges participates in making a master list.

1. Master List Formats

a. Two Category List

In this conventional form of a master list, the evaluator is required to assign every message to one of two categories "relevant" or "not relevant," where relevance is with respect to a given written query. We have found this form of list to be appropriate only under two extreme conditions: (i) when the collection contains only material peripherally relevant to the query, in which case the distinction is really between "peripherally relevant" and "not relevant"; or (ii) when the collection contains a good number of highly relevant messages, in which case the distinction is really between "highly relevant" and "everything else." Thus, what is meant by "relevant" may vary widely from one query to another. Further commentary on two-level evaluations are found in Sections II and IV. We have used this form of a master list for certain studies. We have found, however, that it tends to be too gross to be useful for certain other applications; for example, comparative evaluation of associative retrieval search options given a full-text query.

b. N-Category List

The evaluator is required to assign every message to one of N categories of relevance which are specified in advance. In several of our tests, for example, we have used five categories: (4) Highly Relevant, (3) Moderately Relevant, (2) Nominally Relevant, (1) Peripherally Relevant, and (0) Not Relevant. We have found that this form of list is relatively easy to compile, and is well suited to an important range of questions which turn up only one or two highly relevant messages, but a large number of partially relevant ones. To be useful for later analysis, we have found that it is desirable to attach discrete numerical value scores to each category.

c. Ranked List

The evaluator does not assign messages to relevance categories, but is asked to sort them into decreasing order of relevance. This procedure makes all judgments relative to one another. To be realistic, the evaluator must be allowed to group many messages together into "tie" categories, for he may often have no basis on which to rank one before another. This particularly holds for low-relevance messages. The main problem with a ranked list is that the distance between ranks is not controlled and left unspecified. For example, if messages are ranked A, B, C, D, E, it could be that A is much better than B and B is just slightly better than C, or perhaps A and B are nearly tie and both much better than C, etc.

d. Numerically Scored List

The evaluator is asked to assign each message a relevance value on some continuous scale. The principal difficulty is simply that the evaluator may not be willing or able to distinguish fine fractional relevance values. In principle, all messages could end up with different values, although in practice an evaluator may wish to use only a few discrete values, allowing many ties.

In most of our tests involving only a single evaluator, we have found that the most useful form of a master list is an N-category list (b) with discrete numerical scores attached to each category. Note that this is the same as a numerically scored list (d) with only certain discrete scores being allowed. We have typically used five categories. For certain special applications, we

have used only two. The reason for this choice over a continuous-scale numerically scored list was in part to provide interpretations of scores for the evaluator (i.e., "4" is highly relevant, "3" is moderately relevant, etc.), in part to simplify subsequent processing of evaluator data.

2. Multi-Evaluators

The reason for considering use of multiple evaluators is because one may doubt the "representativeness" of a single evaluator and question the adequacy of his relevance judgments. There appear to be two main avenues open to making a master list when multiple evaluators are involved; they are:

a. Negotiation and Consensus

The evaluators work together as a group and discuss and argue the pros and cons of each relevance judgment until they agree on a decision. Working in this way, they may prepare a master list in any of the forms already described. The difficulties with this method are that there are usually enough disagreements to generate a great many discussions and possible arguments, so the process tends to be very slow and inefficient exercise in group dynamics. A possible alternative peril is that one judge might be more aggressive and vociferous than the others and tend to overwhelm them systematically.

b. Combining Individual Lists

The evaluators each prepare lists separately and individually, and then the lists are combined in some way to make a single over-all master list. Because disagreements have to be accommodated during the combining, the resultant list usually is of the "Numerically Scored" type, regardless of the forms of the original lists.

The most elementary approach to combining individual lists involves nothing more than averaging scores. If each individual list is of the "Numerically Scores" type, then the master relevance score of a message can be taken to be the average of the scores assigned to it by the individual evaluators. We have investigated additional techniques for combining lists, particularly methods for "unbiasing"

individual scores before averaging them, to take into account the fact that some judges may consistently assign much higher or lower scores than others.

We have developed two general approaches to the creation of an unbiased master list, a rather involved but thorough "Error Matrix" approach, and a much more tractable "Simple Unbiasing" approach. These are described briefly in Appendix A. We have done only a limited amount of work using multi-evaluators -- two independent tests employing three evaluators apiece. However, the limited experience gained there seems to indicate that unbiasing of any kind was not necessary for our purposes -- simple averaging appeared to be all that was necessary to produce an adequate master list. Also, our experience seems to indicate that for many practical applications use of multiple evaluators is unnecessary (see Section VI).

D. Economics of Evaluation

The basic cost unit in the evaluation process is the Unit Relevance Judgment (URJ); i.e., the amount of time and effort spent by one evaluator in assessing the relevance of one message to one given query. Given a set of objectives of an evaluation process, one has to decide (a) how many URJs can one afford, (b) how to allocate URJs commensurately with the objectives, and (c) an experimental design which gets the most URJs for a given amount of evaluator time. Basically, one has to decide on numbers: how many types of queries are going to be processed, and how many queries of each type; how many messages will be evaluated for each query, and to what depth; how many evaluators will be used for each query-message combination. The nature of these decisions can best be illustrated by discussion of a few examples, all of which bear on our experimental work discussed in Section VI.

Example 1

The first example is designed to illustrate the necessity for economic compromise in deciding on an evaluation strategy and is in the nature of a reductio ad absurdum. Suppose that one wished thoroughly to evaluate six different searching strategies and made the following decisions regarding experimental design.

Number of Search Strategies	6
Number Types of Queries to be Tested	4
Number of Queries of Each Type	10
Number of Judges per Query	10
Sample Size Evaluated per Query (Messages)	400
Amount of Time per URJ	1-1/2 minutes*

The numbers all look eminently reasonable. However, a slight amount of slide rule manipulation indicates that, if each strategy is checked out against all 40 queries, and if each output is evaluated by all 10 judges looking at 400 messages each, about 13-1/2 man years of evaluator effort would be required -- assuming a 40-hour nonstop work week, normal vacation, etc. We need not comment on the corresponding cost, nor would we consider such a brute force approach. The main point is that decisions as to experimental design must be made in the light of cost of evaluator effort and that a naive all-purpose exhaustive evaluation effort can be impossibly expensive.

Example 2

As a second example, suppose that an objective is to determine whether multiple evaluators are necessary for evaluating certain material, given that they have comparable initial understanding of queries of a given type. Suppose, moreover, that there are other competing objectives so that only a relatively minor investment of effort could be devoted to a preliminary appraisal of this problem. We faced this particular situation and reasoned that the most important question in this case is "how consistent are evaluators' judgments?" and elected to proceed as follows:

* We estimate that this is about the amount of time required per URJ for our collection, counting setup and double-checking of doubtful cases.

Number of Search Strategies	2
Number of Types of Queries	1
Number of Queries	2
Number of Judges	3
Sample Size (Messages)	300
Amount of Time per URJ	1-1/2 minutes

The resultant requirement was for about 2-1/2 man weeks of evaluator effort, a reasonable amount; and some of the URJ evaluations were applicable to other experiments. This particular experiment was designed to answer but one question: given a series of retrieval performance characteristic curves all produced by one judge J_1 , can these be trusted? Specifically, are these curves representative of the ones which would be produced by a panel of three judges J_1 , J_2 , and J_3 ? The test shed considerable light on this question, but gave little information about effects of differing search strategies, queries, or types of queries.

Example 3

In this third example, suppose that the objective is to investigate the effects of differing search strategies on a specific type of query. We conducted some tests with:

Number of Search Strategies	6
Number of Types of Queries	1
Number of Queries	4
Number of Judges	1
Sample Size	400
Time per URJ	1-1/2 minutes

The requirement here was for about six man weeks of evaluator effort, some of which could be overlapped with the experiment in the previous example. Because of the positive outcome of the experiment described in Example 2, we felt comfortable in using only one evaluator for this experiment.

We could go on to give several additional examples, but our main point is clear: the evaluation process must be logically structured into specific steps and specific experiments appropriate to the desired objectives. The economics of testing are formidable in the absence of such a structure.

Our own work has been mainly oriented towards explorations of system alternatives and gaining better understanding of what retrieval systems do and what evaluation consists of. Our effort has been allocated in this direction; and, as a result, there are several possible evaluation objectives we have not been able to address. For example, we have not attempted to compare the effectiveness of a fully automatic associative system under conditions of automatic indexing against that of a manually-indexed coordinate retrieval system using logical formulas and a human system intermediary.

E. Insight-Oriented vs. Proof-Oriented Tests

We have just discussed how economic considerations require that choices be made as to where and how available evaluator manpower be invested. One decision which has to be made is whether the effort is to be put into proof-oriented experiments or into insight-oriented experiments. By the former, we mean concentration of evaluator effort into large-scale systematic and statistically valid investigations of one type of variability with other conditions being relatively fixed; by the latter we mean a spreading of evaluator effort over a number of investigative forays designed to give not proof but insight as to how the experimental variables interact with one another.

A proof-oriented experiment should lead to a well-defined statement of conclusion backed up with an analysis of variance of the results and identified confidence limits. However, such experiments are premature unless one knows exactly what one wants to prove and the conditions under which the proof is interesting. It is not of much interest to know that search option A is proven to be better than search option B under given conditions with confidence 0.9999 -- what is really of interest is whether the given conditions are actually realistic, how much better is A than B, and is this better enough to be of real concern?

Insight-oriented experiments may or may not lead to well-defined conclusions, and one such experiment may or may not be sufficiently meaningful statistically to constitute convincing proof in the face of withering doubt. However, several such tests can sometimes be performed for the cost of one proof-oriented test, and the pattern of observed results might tell a lot more about the system being investigated than any single test, no matter how firm the conclusions of that one test are.

Given our own objectives of exploratory laboratory evaluation, we have found it most fruitful to invest in insight-oriented tests. However, proof-oriented tests are clearly desirable when specific working systems are to be evaluated within given organizational environments, and decisions as to major investments in operational retrieval hardware/software must be made.

The foregoing notwithstanding, we have been alert to the need for rigor at key junctures in our work. In particular, one specific and highly important line of experimentation, described in Subsection A of Section VI, depends on a particular sampling strategy. An analysis of the statistical reliability of this strategy is given in Appendix D, and is typical of the kind of mathematical analyses which must be gone through when proof-oriented testing is involved.

SECTION IV
MEASURES AND TOOLS FOR EVALUATION AND COMPARISON
OF INFORMATION RETRIEVAL SYSTEMS*

A. Introduction

In this section we are concerned with formal methodologies for representing, analyzing, and summarizing the experimental data obtained for measuring a retrieval system's ability to respond to queries. The measurement aspects of experimental design are brought into sharper focus as the problems of comparing system performance with a standard or with the performance of other systems are discussed; other aspects of the comparison problem (how to establish a standard using several judges, the contribution of statistical sampling procedures, etc.) are treated in other sections. Emphasis here is on the situation that arises when the system's responses to a class of queries are given, when the corresponding standard (ideal, master) responses are also given, and the problem of comparing observed responses with the preferred responses requires attention.

Just as different information retrieval systems produce differing responses to the same query, the various search options within a given system also yield up differing outputs. We have found it simplest to regard each distinct option as a separate "system." The problem at hand is to be able to analyze the responses of a system, to compare these responses with the responses of competing systems, and to evaluate them in terms of some absolute measure. Some of the procedures used in the past for these purposes are reviewed in the light of our objectives, and some interesting new alternatives we have investigated are presented. Finally, we explain the graphical summary charts which we have found to be most useful and revealing in our experimental work and identify some of the attributes which more advanced measures and measuring methods should possess to be useful in future work.

B. Definitions

Because this section deals with a restricted (though central) situation encountered in evaluation work, certain terms that are almost self-definitive, or have been discussed at length elsewhere, are nonetheless used in a specialized technical sense within the scope of this discussion.

* The contents of this section as well as Appendices A and B are in part originally due to S. Pollock.

1. Messages

The messages are packages of text which communicate information when interpreted; they have the general characteristics discussed in Section II.

2. Collections

The collection is the set of all stored messages. We assume that they are all potentially available to the information seeker.

3. Query

A query is a statement, recorded in English, of the information that is sought; it is a verbalized expression of an underlying need for information. It may be verbalized in the form of a direct question or in such other forms as "Tell me about ...", as discussed earlier. We assume that the query is the most complete initial basis available for determining the requestor's information need.

4. Search Formulation

A search formulation is the input to the machine portion of a retrieval system. It may consist of the original text of a query, or of key words, or of some other encoding of the query depending on the human steps of preprocessing employed.

5. Relevance Values

Relevance is an extremely subjective element in any analysis of this type. Nevertheless, we stipulate that, within a pattern of conventions that control assignment and interpretation, numerical values may be placed upon each of the messages in the collection which indicate the relevance of the message to a designated query. As discussed in Sections II and III, values may be assigned by a particular "judge" or a "panel of judges" or a value might be estimated by a machine procedure. But, in general, it always is extremely dependent upon the assigning agency and upon the specific query.

The units for this value may be established by whatever conventions are adopted, although it is convenient in the presentation of this chapter (without loss of generality) to measure values on a 0 to 1 scale, with 1 representing the perfect degree of satisfaction, 0 being the perfect degree of dissatisfaction. For example, consider the query "Tell me about XYZ." The message in the collection that is demonstrably "Everything there is to know about XYZ" might be assigned (e.g., by the

requestor) a value close to (if not exactly) 1, while a message that treats "Nothing about XYZ but all about ABC" might have an assigned value close to 0.

6. Master Values

These are hypothesized perfect or true values of the relevance of messages. While relevance values as in (5) can be assigned by any agency -- machine or human -- the master values are inherent in each message-query pair; they can be thought of as assigned by perfect judges, and they are not related to any particular information retrieval scheme. As discussed in Section II, the existence of such values is clearly a philosophical question. We assume that they can be approximated in practice by a suitably selected human individual or jury.

7. System

A system is a general means of searching through the collection to obtain information concerning messages. Typically, a system is directed to identify the messages that are relevant to a particular query. The search options mentioned in Section I and discussed in Section VI are examples of systems.

8. Output List

This is the name given to the output obtained when a system is given a search formulation and applies it to the collection. It is a listing of some or all of the messages in the collection, together with appended information interpretable as the system's estimates of the relevance of each of the listed messages.

The list may be, for example, one of the forms:

- (a) Messages A, B, C, F, G are responsive to the query, the rest are not.
- (b) Messages A, B, C are very responsive, messages F, G, H are of interest, the rest are not.
- (c) In order of satisfaction of the query, the messages are A, C, F, B, H, G, K, L, ..., etc.
- (d) The value of each message, with respect to the query, is

Message	A	C	F	B	H	G	K	L
Value	0.9	0.8	0.4	0.3	0.2	0.05	0.05	0.03

- (e) Combinations of the above. Although retrieval systems exist with each of these forms of output, we are mostly concerned with systems with type (d) output. The associative retrieval systems, for example, fall in this category.

9. Master List

The master list* is a listing of all messages in a sample selected for evaluation, with the master values observably affixed to each message. For a particular query, it is often convenient to think of the messages in the master list as being ordered in decreasing master value.

10. Recall Ratio

This measure is mainly useful when the values assigned to the messages by the system are either 0 or 1 and, in fact, the master values are also either 0 or 1. In this case, the recall ratio is the fraction of messages assigned 1 that have master value 1, divided by the total number of messages in the collection that have master value 1.

11. Precision Ratio

This ratio again has meaning only when values can be either 0 or 1. In this case, the precision ratio is the fraction of messages that have been assigned value 1 that have master value 1.

C. Previous Statistics and Measures

Several measures have been devised to compare and evaluate information retrieval systems by analysis of the statistics of the lists produced by competing systems. Some analyses appearing in the literature have compared lists between two competing systems rather than the list of a system with a master list. There are questions concerning the applicability of these statistics, however, which will be brought up after a brief summary of their definitions and characteristics.

* For further discussion of the master list, see Section III, Subsection C and Appendix A.

1. Precision and Recall Ratios

Easily the most widely-used summary statistics are the Precision and Recall Ratios.* These measures, as defined earlier, are absolute measures in that a master list must be specified for determining the true (master) value of each message, specifically 0 (not relevant) or 1 (relevant). Precision and recall ratio are defined over all unordered subsets of messages. The higher these ratios for retrieved subsets, the better the system is interpreted to be. It has been observed (21) that these measures are not completely independent; considerable attention has also been paid to the observation that as one ratio goes up, there is a tendency for the other to go down (22). For example, a retrieval system that assigns the value 1 to all messages in the collection will in doing so assign value 1 to all messages that have master value 1, and so the recall ratio will be unity. However, the precision ratio will then fall to the fraction of messages in the collection that have master value 1.

A probabilistic interpretation of these ratios is often helpful. If a class of queries is given over which the recall and precision ratios are well-behaved statistically, we may define for any message-query pair the relations:

r = recall ratio = $\text{prob. } \{R/S\}$

p = precision ratio = $\text{prob. } \{S/R\}$

where the statements $\{R\}$ and $\{S\}$ are defined

R = message is assigned value 1 by the system (is "Retrieved")

S = message has master value 1 ("Satisfies the query")

If L is the number of messages in the collection and L_1 is the number of messages in the collection with master value 1, then the following conditional probability statement holds:

* These statistics have received such widespread use as almost to preclude employment of others. In particular, see the various writings of Cleverdon et al (7)-(10), Richmond (11), Swanson (12), Taube (3), Hillman (4), and one of our previous reports (20).

$$p = \text{prob. } \{S/R\} = \frac{\text{prob. } \{R/S\} \text{ prob. } \{S\}}{\text{prob. } \{R/S\} \text{ prob. } \{S\} + \text{prob. } \{R/\bar{S}\} \text{ prob. } \{\bar{S}\}}$$

$$p = \frac{r \frac{L_1}{L}}{r \frac{L_1}{L} + f \left(1 - \frac{L_1}{L}\right)} = \frac{1}{1 + \frac{f(L-L_1)}{r L_1}} \quad (1)$$

where $f = \text{prob. } \{R/\bar{S}\} = \text{False relevance assignment probability (i.e., probability of assigning value 1 when the master value is 0)}$

These measures are closely related to the conventional error measures:

Type 1 error = $\alpha = \text{prob. } \{\bar{R}/S\}$

Type 2 error = $\beta = f = \text{prob. } \{R/\bar{S}\}$

Note that $\alpha = 1 - r$.

2. Salton's Normalized Precision and Recall*

Certain summary statistics Salton has used also assume that the master values are 0 or 1, but modifications are made for the case when the output of the retrieval system is a ranked list. Salton defines these measures as

$$R_{\text{norm}} = \frac{1}{L} \sum_{j=1}^L r_j, \quad P_{\text{norm}} = \frac{1}{L} \sum_{j=1}^L p_j$$

where r_j and p_j are the recall and precision for the set consisting of the messages ranked 1 through j by the system. These measures are averages of the precision and recall over all possible sizes of the document set one could treat as "retrieved."

* See (13).

Salton indicates that these measures can be written

$$R_{\text{norm}} = 1 - \frac{\sum_{i=1}^{L_1} s_i \sum_{i=1}^{L_1} i}{L_1 (L - L_1)}$$

$$P_{\text{norm}} = 1 - \frac{\sum_{i=1}^{L_1} \ln s_i - \sum_{i=1}^{L_1} \ln i}{\ln \frac{L!}{L_1! (L - L_1)!}}$$

which are forms more suitable to computation. In these approximation formulas, s_i is the rank (in decreasing correlation order with the search request) of the i th relevant document in the collection.

3. Rank Order Statistics

These statistics have their origin in classical problems involving the comparison of various rankings. The statistics can be used in both an absolute way and a relative way. That is, two lists may be compared against each other, or individual lists may be compared against master lists.

Most commonly used rank-order statistics are specific cases of a general rank-correlation coefficient.* This coefficient is defined as follows.

Consider the two lists of numbers:

list A: $x_1, x_2, x_3, \dots, x_n$

list B: $y_1, y_2, y_3, \dots, y_n$

These are the lists we wish to compare. The methodology permits us to express, for each list, how much "it matters" to find x_i in

* See Kendall, M. G., Rank Correlation Methods, (23).

position i and x_j in position j . We define for the A list, the score a_{ij} for each pair of numbers (x_i, x_j) , $j = 1, 2, \dots, n$. A similar score b_{ij} is defined for the B list. The only limitation on these scores is the symmetry requirements.

$$a_{ij} = -a_{ji}, b_{ij} = -b_{ji} \text{ (so that } a_{ii} = b_{ii} = 0 \text{)}.$$

A general rank-correlation coefficient Γ is then defined to be

$$\Gamma = \frac{\sum a_{ij} b_{ij}}{\left[\sum a_{ij}^2 \sum b_{ij}^2 \right]^{1/2}} \quad (2)$$

where the \sum indicates summation over all $i, j = 1, 2, 3, \dots, n$. Kendall shows that when $a_{ij} = x_j - x_i$ and $b_{ij} = y_j - y_i$, then Γ becomes the ordinary product-moment correlation of x and y . Note that unlike the precision and recall conventions, a single number results from use of (2).

a. Kendall's τ Statistic

The simplest and most widely-used form of this coefficient is Kendall's τ statistic. Γ is equal to τ when a_{ij} and b_{ij} are defined

$$a_{ij} = \begin{cases} +1 & p_i < p_j \\ 0 & p_i = p_j \\ -1 & p_i > p_j \end{cases}$$

$$b_{ij} = \begin{cases} +1 & q_i < q_j \\ 0 & q_i = q_j \\ -1 & q_i > q_j \end{cases}$$

where p_i and q_i are the rank of the i th element of the A and B lists. This τ statistic exhibits the properties

- (1) $\tau = 1$ when the two lists are identical;
- (2) $\tau = -1$ when one list is exactly the inverse of the other;
- (3) The expected value of τ is zero when both lists are random arrangements of the elements x_i and y_i .

It also may be shown that this statistic is a linear function of the number of pairwise interchanges of neighboring elements (x_i, x_{i+1}) required to change the A list into the B list (or vice versa).

b. Spearman's ρ Statistic

This coefficient is obtained by assigning:

$$a_{ij} = p_i - p_j$$

$$b_{ij} = q_i - q_j ,$$

the differences in rank of the i th and j th elements of the lists. Spearman's ρ may also be shown to have the properties listed above for the τ statistic. In addition, it gives more weight than the τ statistic to the differences in the rank of elements of the list, as they get further apart.

4. The M-V Statistic

A modification of Kendall's τ has been developed by Kay Mazuy and Emilio Venezian of our staff, called the M-V statistic.* This statistic is related to Kendall's τ and is defined by letting:

$$a_{ij} = \begin{cases} +1 & p'_i < p'_j \\ 0 & p'_i = p'_j \\ -1 & p'_i > p'_j \end{cases}$$

* Mathematical properties of these measures have been described elsewhere; see references (1) and (24).

where p'_i is a modified ranking of the A list. This modified ranking is such that only the first M elements of the list are arranged according to the original ranking, and the remaining N-M elements are grouped together into an (N-M)-fold tie with the same last rank. Thus

$$p'_i = \min (p_i, M+1) .$$

The same holds for b_{ij} and the B list. This statistic, call it $\theta(M)$, has introduced a free variable M -- the length of the list (from the beginning) that will be used in computing the rank-order coefficient. The advantage of having such a free variable in the measure of effectiveness of systems will be discussed later on. It suffices at this point to note that the $\theta(M)$ of a particular pair of lists might vary considerably with M.

D. The Disadvantages of Previous Statistics

1. Precision and Recall Ratios

These ratios are useful for certain applications when relevance can be assessed on a black or white basis and when the search logic groups all retrieved documents together in an unordered batch. However, for many applications they have the primary disadvantage that they are too gross to measure the system properties of most interest. First of all, both assigned and master values must be either 0 or 1. That is, a message must be said to be either entirely relevant to the query or not at all relevant. No spectrum of relevance in between is allowed.

Our experience with our own message collection has been that for a typical highly specific request there are exceedingly few if any messages with very high relevance but a substantial number with varying degrees of partial relevance. Under these conditions, making relevance decisions on a strictly 0 or 1 basis is unnatural and often results in quite arbitrary decisions. A second problem with precision and recall is that all messages are regarded to be grouped into two classes, "retrieved" and "not retrieved"; there is no way of measuring the effectiveness of a machine-generated ranking of retrieved messages. These difficulties hold also for the Type 1 and Type 2 error measures.

2. Salton's Normalized Precision and Recall

These measures overcome one of the difficulties with precision and recall -- namely, they can be used with ranked machine output lists. They still continue to be too gross for certain purposes, however. One reason for this is that the master values must continue to be 0 or 1

and nothing in between. Another reason is that the normalized precision and recall measures do not at all convey essential information about the retrieval characteristics of the system under study because the entire list of retrieved messages is used as a basis for comparison. For example, two systems having identical R_{norm} and P_{norm} values could be such that the first is far better if one wants to look at only five to ten messages, but the second is far better than the first if one is willing to look at a list of one or two hundred messages.

3. Kendall's τ

The major disadvantage with this statistic is also the fact that the entire list is used as a basis for comparison. In other words, an interchange of elements with master rank 3 and 7, for example, in the assigned list, will be just as important as interchange of ranks 103 and 107. However, users of the system would tend to regard the first interchange as being much more serious than the second. Clearly, a realistic measure of the list produced by a retrieval system should have some means of weighting appropriately the more "important" initial parts of the list.

4. The M-V Statistic

The M-V statistic was created to do such a weighting. This statistic counts only the first M elements of the list, puts the rest in a tie for the (M+1)th place, and then performs the Kendall arithmetic; it has two disadvantages. First of all, once M has been selected, the process essentially weights interchanges among the first M members of the list equally. Secondly, the calculations are cumbersome.

5. Summary

Perhaps the most glaring fault with all the statistics mentioned above is that they treat the lists produced by retrieval systems in a fairly abstract way. That is, they go too far in boiling the data in the lists down to one or two numbers, and sometimes these numbers seem to bear little relationship to the use to which the lists must be directed. Our experimental work discussed elsewhere in this report has brought this fact home to us with great clarity and has led us sometimes to prefer use of some simple but very effective graphical techniques which are discussed in Subsection F below.

E. Toward a Rational Measure of Effectiveness

This section discusses characteristics desirable in a realistic measure of a retrieval system's effectiveness. As exhibited by the M-V modification of Kendall's τ statistic, there is a desirability to modify available statistics to produce something which can be considered to be a more "useful" or more "applicable" measure of the retrieval capabilities of a particular system (as represented by the list it produces). Perhaps the most straightforward way of looking at this problem is to ask about the end purpose of these lists: what will be done with them? How will they be used?

1. The Officer

Consider a requestor who wishes to find only one message of relevance and is only willing to read two or three messages to find it. Let us call him the "officer." To evaluate a system's response with respect to his rather extreme criterion, the chosen measure should give a high score to a retrieval system that places only one, two or three messages before the officer and insures that these messages have high probability of relevance. The officer is uninterested in exhaustive search and is indifferent to the system's failure to present all relevant messages. He does not care what the probability of relevance of the 50th message retrieved is. He is just interested in finding one and refuses to read beyond three. The system's performance is judged by the first three messages presented, and the system is penalized for "top-of-the-list" failures even though it might have a perfectly good list of relevant messages to follow.

2. The Analyst

It is possible for a requestor to be interested in compiling, say, a 100-message bibliography (responsive to some query), which is to be studied and analyzed at leisure. Let us call him the "analyst." The analyst might be content with a list of perhaps 400 messages from which he can select 100 particularly pertinent messages; he is probably not seriously concerned about the order in which these 400 are presented to him. He does not mind at all if the first three are not relevant if a high proportion of the rest are. On the other hand, the analyst is not impressed by a system that only does well for the first two or three messages and then begins to produce irrelevant material.

3. The Browser

Another possible user of the retrieval system's output -- the "browser" -- might be interested in a sequential examination and exploration. When the browser is given a list of relevant messages, he goes through them; in order of computed value if a ranking is given, or

just randomly if an unranked set is given. The browser proceeds to look through the messages and will stop when he feels that he has obtained enough information or has satisfied himself as to the amount of information that he wants. From the browser's viewpoint, good retrieval performance means that at each stage in the sequential process, probability of relevance of the next item is maximal.

The many possible uses of a list of retrieved messages suggest that a desirable measure of retrieval effectiveness should be responsive to these variable needs. In particular, there should be a "depth of search parameter" available to be varied such that if the parameter is at one end of the scale, top-of-the-list performance for only one or two messages is emphasized; if the parameter is at the other end of the scale, the whole collection becomes important, perhaps with ranking of the collection relatively unimportant. With the parameter in the middle of the scale, it should be a mixture of the two. The M-V statistic comes close to doing this, and other schemes might also be appropriate for the same objective. These are discussed in Subsection G. We turn first, however, to a discussion of the graphical techniques which we have found to be extremely useful in practice, and for many applications to largely obviate the need for any formal performance measure.

F. Performance Characteristic Curves

Our experience has been that so long as a retrieval system is likely to be used on different occasions with different search objectives in mind, the system's performance is more meaningfully characterized by curves or by families of curves than by numerical measures which assume or otherwise incorporate a standard objective. A retrieval system is like a vacuum tube or a transistor in this regard -- a curve explains the interrelationship between two variables more effectively than does a table of numbers, and families of curves can be used effectively to show how three or more variables interact.

The performance characteristic curves we have found to be most useful exhibit plots of cumulative value of retrieved messages as a function of rank. The value of a retrieved message is that assigned by the master list. A single diagram can contain curves for a single query or for average performance over several queries, and can show performance for several different systems or search options in juxtaposition.

Some features which can be observed using performance characteristics curves are illustrated in the graphs of Tables IV-1 and IV-2. Table IV-1 exhibits curves for three hypothetical search options, called B, C, and D. The curves are supposed to represent average observed performance for these options, assuming a given general query type. In this simplified example, it is assumed that the master list contains only 0 and 1 values and that on the average for the queries studied, a

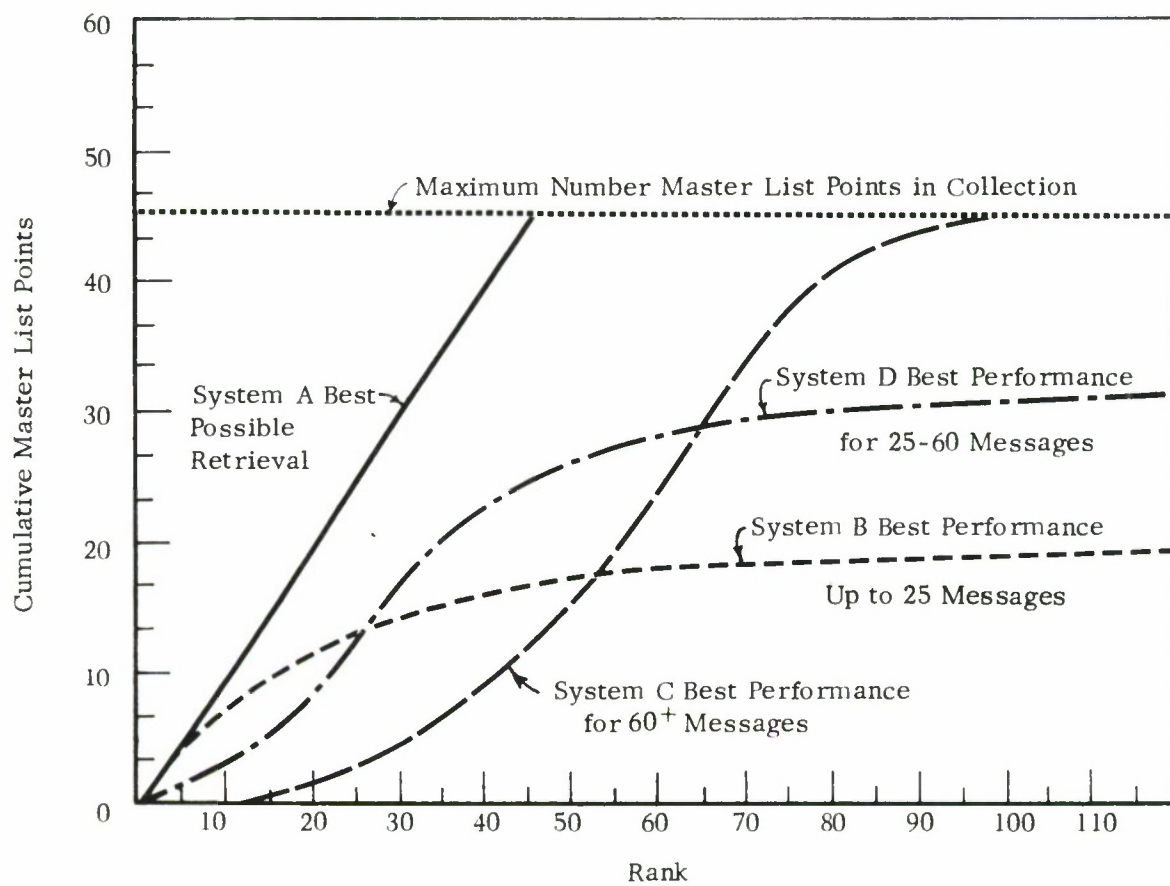


TABLE IV-1 PERFORMANCE CHARACTERISTIC CURVES FOR THREE HYPOTHETICAL SEARCH OPTIONS

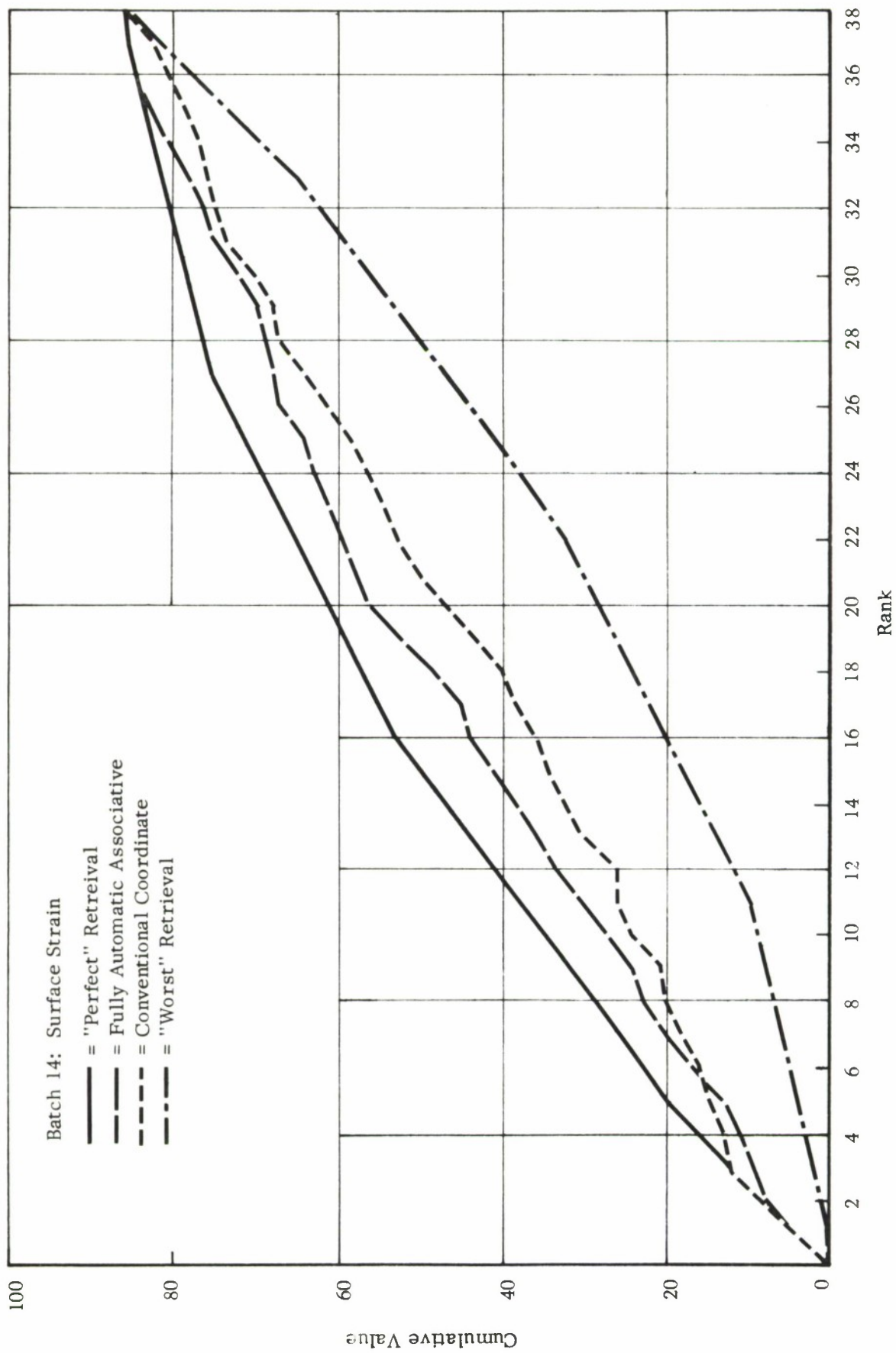


TABLE IV-2 PERFORMANCE CHARACTERISTIC CURVES FOR THE SUBJECT HEADING QUERY "SURFACE STRAIN"

total of 45 messages in the master list have value 1; best possible retrieval performance would therefore be represented by curve A, which simply means that all relevant messages are retrieved before any others.

Interpretation of the table shows at once that system B is the best performer as long as less than 25 messages are retrieved and, indeed, works very well for less than ten messages. It is clearly the search option that the officer wants to use. On the other hand, system B is not very good for the analyst, who is willing to look at up to hundreds of messages. The analyst is better off using option C which does not give a very good head start but eventually yields the most relevant messages. Finally, the browser might well prefer option D, because it gives him the best results over the range of 25-45 messages. A clever browser might wish to mix strategies, say, by switching from system B to system D after the first 25 messages; however, he would have then to start at the bottom of the system D curve and possibly have to re-read messages.

As a second example, Table IV-2 is the retrieval characteristic curve for an actual request, Surface Strain, and shows a comparison of two search options made based on the use of the Fully-Automatic Associative search option and of the Conventional Coordinate retrieval option on the same material. The master list for this query contains scores which vary between 4 (very relevant) and 0 (not relevant). There are a total of 85 relevance points in the sample studied. The uppermost curve is for "perfect" retrieval of messages in the sample and shows what would happen if the machine retrieved all the messages with value 4 first (there happen to be five of them) followed next by retrieving the 11 messages with value 3, etc. The lowermost curve shows what would happen if the messages within the sample were arranged in the worst order; that is, first the messages in the sample with value 0, then those with value 1, etc. Curves for what the machine actually did for the two options studied fall in between. As might be expected, the Conventional Coordinate (CC) option tends to produce a curve about halfway between the two extremes; the Fully-Automatic Associative (FAA) option produces a somewhat better ordering. However, considering only the first three messages, the CC option is slightly better for this particular query. The real advantage of the FAA option is only in the midrange between twelve and 26 messages, and at 38 messages the two options are equal again.

Such over-all features of retrieval performance are exceedingly difficult to capture and compare meaningfully by means of summary measures, but they become quite obvious when displayed by means of performance characteristic curves.

G. Measures of Performance Features

In our own experimental evaluation work, we have not found it necessary so far to go beyond the use of various kinds of performance characteristic graphs as a basis for making comparative judgments among retrieval systems. Nonetheless, we have been concerned with the

development of more adequate evaluation statistics, ones which take the user's depth-of-search requirements into account. We have identified four such statistics, and the formulas for these are derived and presented in Appendix B. They are:

1. A Normalized Sliding Ratio Statistic

The conceptual basis for this statistic is a generalization of the precision and recall ratios. These ratios are generalized to enable consideration of a master list with a spectrum of possible values, not just 0 or 1, and are expressed as functions of j , the number of messages retrieved. A single ratio -- for performance measurement -- $\mu(j)$ is then formed which has several interesting properties. When plotted graphically, $\mu(j)$ yields a normalized performance characteristics curve with all values between 0 and 1. Using these normalized curves, it is possible to compare performance for different queries using different master lists. The measure can be modified to suit different depth-of-search requirements by choosing different values of j . For example, to satisfy the needs of the officer who wants to look at only three or fewer messages, evaluation could be based on $\mu(3)$. To satisfy the analyst who is willing to look at 400 messages, evaluation could be based on $\mu(400)$. Other users with intermediate depth-of-search requirements might wish to use other values of j when using $\mu(j)$ to evaluate.

2. A Browser's Statistic

This statistic is designed to meet the need of the browser as we have characterized him previously. Specifically, we suppose that the value of a message is related directly to the probability that that particular message will completely satisfy the browser's curiosity, thus ending the need for further message inspection. This being the assumed case, then we consider using, as a measure of effectiveness, J , the expected number of messages it takes for a browser to be satisfied, given a particular ranked list.

3. A Weighted Rank-Correlation Statistic

The M-V statistic is essentially a technique to weigh only certain portions of the retrieved list. The portion up to rank R is weighted unity, the portion past R is weighted zero. A more mathematically tractable approach, and one more in accord with our intuitive notion of variable depth-of-search requirements, is to weight the whole list with decreasing values from the head of the list to the end. The weighting function which first comes to mind and seems most natural is simple exponential weighting. The formula we develop contains a constant α with possible values between 0 and 1. When we set $\alpha = 1$, the measure simply becomes Kendall's τ , a measure well-suited for the

analyst for whom the ordering is of equal importance throughout the list. When we set $\alpha = 0$, only the first message in the list is counted, and the measure is more suitable for the officer. To evaluate for depth-of-search requirements intermediate between that of the officer and that of the analyst, intermediate values of α can be used.

4. Cost-Matrix Measures

Not only may one user's depth-of-search requirements be quite different from that of another, but a specific user may wish to vary his criteria for expressing satisfaction with a search, depending on his needs of the moment. A way of allowing each user to reflect his own depth-of-search criteria in a single evaluation measure is to allow him to specify a "cost" matrix. This might be of the form C_{ij} where C_{ij} represents the cost of having a message, whose master ranking is i , assigned the rank j by the system. Using such a matrix, the user can assign cost penalties to each such assignment independent of other assignments, and the value of a given list of assignments may be obtained from the matrix. The procedure resembled Kendall's general rank-correlation methodology, and in Appendix B, we indicate how measures can be defined on such matrices to give estimates of over-all retrieval effectiveness.

II. Modeling Retrieval System Performance

Given a specific evaluation measure and a specific retrieval system, it may be possible to construct a parametric mathematical model of how that system is apt to behave with respect to those measures. If one is successful in creating such a model, it may be possible to make observations of how a system works under certain test conditions and then use the model to predict how well it will probably work under other test conditions. In the course of previous work, we have constructed such a model for evaluation of performance of coordinate information retrieval schemes, based on the use of measures related to precision and recall.* This model was used to predict precision and recall performance of these schemes as a function of collection size. So far, we have regarded it as premature to construct such a general evaluation model for associative retrieval systems. Such a model is bound to be fairly complicated because the associative schemes are much more complex than the coordinate ones and require application of more sophisticated measures than precision and recall. Also, up until recently there has been almost no data available on which to base a model. However, we now view the prospect of development of such a general model to be of decided interest and potential value.

* See the Appendix to the Second Edition of Centralization and Documentation (20).

I. Summary

We have reviewed past procedures for measuring performance of retrieval systems with respect to their capabilities for retrieving relevant information and have found that the measures which have been used previously are too gross to be useful for some aspects of evaluation of an association-based system. We have identified a spectrum of different evaluation criteria depending on user depth-of-search requirements. Our experience indicates that comparison of total system performance simultaneously with respect to several of these criteria can be conducted readily using simple performance characteristics curves. These curves can be normalized to enable simultaneous comparison for several different queries, search options, or systems. We suggest that families of these curves provide natural means for describing system performance. Further summary measures and comparison procedures can be based on these curves -- measures which account for differing depth-of-search requirements -- and we have formulated several of these.

SECTION V

EXPERIMENTAL DATA BASES AND RETRIEVAL TOOLS

This section is devoted to (a) brief descriptions of our experimental message collections and a discussion of some of their statistical properties, (b) some interesting comparisons between the manual- and machine-indexing vocabularies and the results of indexing messages with them, and (c) brief descriptions of our retrieval computer programs and other searching tools. The reader mainly interested in measurement of retrieval performance may wish to skip over Subsections A and B of this section on first reading, since these subsections tend to be highly technical.

A. Data Bases

1. The Parent GE Collection

We have done various studies and tests on three distinct message collections, called GE-0, GE-1, and GE-2. The GE-0 collection contains entries for some 70,000 documents, and was in use as the data base of an operational document retrieval system at the time when we purchased rights to use it from a branch of the General Electric Company. The second two collections are selected subsets of the first, each about 10,000 items long. The collections treat aerospace topics in general, but with emphasis on flight propulsion technology -- jet and rocket engines in particular. We acquired the collection in early 1963 and, at that time, it represented about eight years accumulation of material. The GE-0 collection was delivered to us in the form of two sets of magnetic tapes. The first of these is an index tape which lists, for each of some 70,000 papers and articles treated, the UNITERM index terms which were assigned to these documents by professional indexers. A second set of tapes consists of short, informative English abstracts -- 30 to 80 words long typically -- of the information contained in about 45,000 of the documents. All abstracts were provided us except those which were under security classifications or which were considered proprietary by the General Electric Company. About 4,824 distinct UNITERMS have been used to index the 70,000 documents.

The GE-0 collection was originally designed and used to aid in the retrieval of full-length documents. The designers of this system apparently viewed the retrieval of abstracts only as an intermediate step in the over-all document retrieval process. Our view and purpose of using the collection and its subcollections for experimentation has been somewhat different; we viewed the abstracts as being informative messages in their own right -- the fact that the messages tell about the contents of longer documents has not concerned us. We have been very much concerned with whether retrieved messages are relevant to a

query; we have not been directly concerned with the much more complex problem of assessing the relevance of documents which are eventually retrieved through processes involving intermediate use of messages. For this reason, much of what we will have to say here in Subsection B about comparing manual and automatic indexing may not be applicable within the context of the original retrieval application at GE.

2. Corpus GE-1 -- Manually-Indexed

This corpus consists of 9,606 surrogates machine-selected out of the larger GE-0 collection of 70,000 surrogates. By surrogate, we mean the set of index terms manually assigned to a message. The surrogates are formed from a total of 1,087 terms. This vocabulary is called GE-1A and is a machine-selected subset of the manual UNITERM index set of 4,824 terms (vocabulary GE-0A) used in the GE-0 parent collection. Details of the machine procedures for the selection of the subcollection of messages and terms have been given elsewhere.* Briefly, these steps were followed: first, about 1,500 highly technical, meaningless or metallurgical terms were dropped from the original GE-0A list, leaving about 3,300 terms. Those remaining were sorted into order of decreasing use frequency and then every third one was selected. This led to the 1,087-term vocabulary GE-1A. All surrogates in GE-0 consisting of seven or more terms of GE-1A were included in the GE-1 collection. Average depth of indexing is about nine terms per GE-1 surrogate. There exist actual messages (i.e., abstracts) for about 4,500 of the surrogates selected this way; the remainder of the surrogates corresponds to classified or proprietary items.

3. Corpus GE-2 -- Automatically-Indexed

This corpus consists of 10,289 messages automatically indexed by 999 terms. The messages consist essentially of every fourth abstract in the GE-0 collection. We make no use of surrogates; each message is an abstract, and the index terms (single words) are selected by machine from the text of the abstracts.** Typical abstracts are exhibited in Table V-1.

We have studied the statistics of language usage within this collection at length, both for retrieval evaluation and language research purposes. Some of the more relevant of these statistics are mentioned here for general reference. The interested reader may find a wealth of additional statistics having to do with word string usages in our Technical Notes, particularly in the supplements to TN CACL-13.

* See (1). Additional details are given in Technical Note CACL-12, (25).

** Details of the machine selection procedures are given in Technical Note CACL-13, (26).

ADL 9942ER 994B G.E. NUMBER 71488
WADC TR-55-491-6 060062+ FORLANO, E J + KRUMWIEDE, D M ET AL

"TEXT"

RESEARCH ON ELEVATED TEMPERATURE RESISTANT
CERAMIC STRUCTURAL ADHESIVES.
STRONG AND RELATIVELY SHOCK-RESISTANT INORGANIC
ADHESIVES DEVELOPED AND IMPROVED BY ADDITION OF
METAL FILLERS AND RECRYSTALLIZABLE MATERIALS TO
THE GLASSY PHASE. AF33(616)6192

ADL 9943ER 994C G.E. NUMBER 71491
ASD TDR-62-24 010162+ RUDKIN, R L + PARKER, W J + JENKINS, R J

"TEXT"

THERMAL DIFFUSIVITY MEASUREMENTS ON METALS AND
CERAMICS AT HIGH TEMPERATURES.
THE ADAPTION OF THE NRDL FLASH METHOD TO THE
MEASUREMENT OF THE THERMAL DIFFUSIVITY OF METALS
AND CERAMICS AT HIGH TEMPERATURES IS DESCRIBED.
MEASUREMENTS OF THE THERMAL DIFFUSIVITY OF ARMCO
IRON, MOLYBDENUM, TITANIUM, ZIRCONIA, AND
ALUMINA HAVE BEEN MADE UP TO 1200, 1300, 1700,
AND 1100C, RESPECTIVELY. MIPR 33(616)61-7

Sample Abstracts From the GE-2 Corpus

TABLE V-1

The indexing vocabulary of the GE-2 collection, known as the GE-2A vocabulary, consists of 999 machine-selected "terms" representing a total of 1,434 singular and plural word forms (e.g., both "temperature" and "temperatures" are instances of a single term). The GE-2A terms consist of the most frequently occurring coalesced singular and plural forms found within the 10,289-message collection, excluding a list of rather trivial "function" words (listed in TN CACL-13, Suppl.). The 10,289 abstracts provide 446,097 words of running text. Roughly 343,000 or 77% of these word occurrences are accounted for by the GE-2A vocabulary and by the function words. If a nonfunction word occurs in singular and plural an aggregate of at least 56 times, it is used as a term in GE-2A. If either the singular or plural of a term occurs in a message, that term is used to index the message.

The automatic indexing procedure produces an average assignment of 16.7 terms per message. The distribution of number of index terms per message is shown in Table V-2, and it appears to be nearly normal.

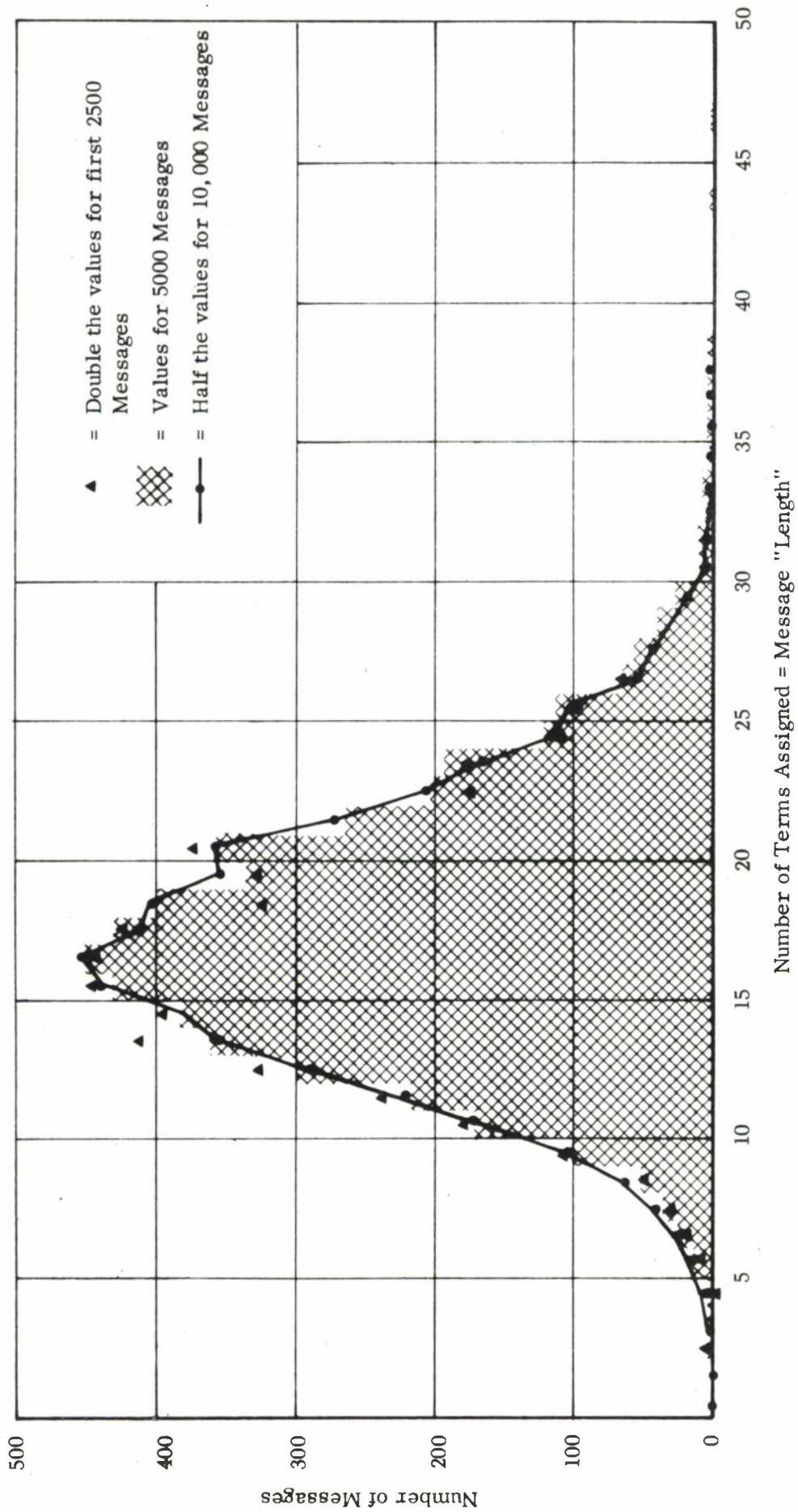


TABLE V-2 DISTRIBUTION OF AUTO-INDEX TERMS PER MESSAGE FOR THE GE-2 COLLECTION AND ITS INITIAL FRAGMENTS

A total of 89.3% of all messages are assigned between 10 and 24 terms. Only 0.5% of the messages have less than six terms assigned to them, and only 0.3% have more than 31 terms assigned. There is a slight historical tendency towards larger assignments of index sets, probably reflecting a trend toward longer and more detailed messages. We do not consider the trend significant. The first 1,000 messages have an average of less than 15 terms per message; the last 1,000 messages have an average of over 17 terms per message (see TN CACL-16, p. 6). Average message length is 44.5 word tokens. Of these, about 20.9 are usages of GE-2A index terms, about 14.1 are instances of function words, and about 9.5 are instances of content words not reflected in the indexing vocabulary.

We have investigated the rank occurrence frequency characteristic for word forms and for context strings of length up to four in the GE-2 collection. The curve for word forms (plotted on a log-log scale) does not exhibit a straight line with slope -1 appropriate to the much discussed Zipf curve.* Instead, it exhibits a distinctive downward-bending bow, such as we have observed in other humanly-assigned indexing vocabularies. Average slope is roughly -1. The curves for context strings of length greater than one are, however, straight lines with varying slopes. The slope for two-word context strings is -0.64, that for three-word context strings is -0.50, and that for four-word context strings is -0.45. The 240 "function" words deleted from eligibility as index terms also turn out to exhibit the Zipf frequency distribution, with slope roughly -1. Actual curves are given in TN CACL-13, Suppl.); some additional statistics of word usage are given in Table V-3.

There are 23,505 distinct word forms occurring in the running 446,097 words of text in the corpus. There are 42,066 distinct word pair strings which occur twice or more in the corpus, 11,619 distinct word triplet strings which occur three times or more, and 3,304 distinct four-word strings which occur three times or more. We have estimated that misspellings account for 42.5% of word forms occurring only once, 16% of forms occurring twice, and 7.3% of forms occurring three times. The average rate of misspelling is one for every other message (see TN CACL-13, Suppl.).

We have performed various informational entropy calculations on the text; because of the general interest in such numbers, the results are summarized here. We found:

* See, for example, Zipf, G. K. Human Behaviour and the Principle of Least Effort. Addison-Wesley, Cambridge (1949).

NATURE OF CORPUS	Complete texts of 10,289 abstracts of documents dealing with aircraft technology (see samples in Table 1).					
LENGTH (# TOKENS)	446,097		NUMBER OF INDEX TERM TYPES		999	
# WORD TYPES	23,505		AVERAGE NO. POSTINGS TERMS/MESSAGE		16.7	
FREQUENCIES OF WORDS MOST USED IN TEXT	of	30,170	on	4,634	high	1,757
	and	16,622	with	3,844	is	1,704
	the	8,911	at	2,770	as	1,689
	in	8,879	by	2,759	heat	1,547
	to	8,045	flow	2,184		
	for	7,406	temperature	2,033		
	a	6,232	from	1,807		
WORDS LEAST USED (i=frequency, n _i = no. of types used i times)	(i)	(n _i)	(i)	(n _i)	(i)	(n _i)
	1	12,485	6	422	11	179
	2	2,929	7	346	12	156
	3	1,370	8	310		
	4	824	9	268		
	5	631	10	221		
EXAMPLES OF TYPES USED ONLY ONCE	aesthetic		affirmed		Africa	
	af 29		afing			
	af 34		afoo			
	afe		afore			
	affaiblissem		afor			
	affectant					
EXAMPLES OF TYPES USED FOUR TIMES	agency		alkalis		aluminizing	
	aggregate		Allen		amenable	
	airflows		allowables		analog	
	airports		allowing		analogies	
EXAMPLES OF INDEX TERMS IN GE-2A INDEX VOCABULARY	ablation		account		action	
	absorption		accuracy		addition	
	abstract		accurate		additive	
	acceleration		acid		adhesive	

Some Statistics of GE-2 Corpus Used for
Associative Retrieval Experiments

TABLE V-3

$$H(x) = - \sum_i p_i \log_2 p_i = 10.36 \pm 0.01 \text{ bits/word form}$$

$$H(x,y) = - \sum_{ij} p_{ij} \log_2 p_{ij} = 16.3 \pm 0.2 \text{ bits/digram}$$

$$H(x,y,z) = - \sum_{ijk} p_{ijk} \log_2 p_{ijk} = 17.9 \pm 0.4 \text{ bits/trigram}$$

The first figure was obtained by direct computation based on exhaustive information on word form frequencies; the last two involved some estimating but are within the error bounds indicated. Using the formula $H_x(Y) = H(x,y) - H(x)$, we obtained

Uncertainty Associated with Random Selection of a Word From Text	$= H(x) = 10.36 \pm 0.01 \text{ bits}$
--	--

Uncertainty Associated with Random Selection of a Word From Text, Knowing the Word in the Previous Position	$= H_x(Y) = 5.9 \pm 0.4 \text{ bits}$
---	---------------------------------------

Uncertainty Associated with Random Selection of a Word from Text, Knowing the Preceding Word Digram	$= H_{xy}(Z) = 1.6 \pm 0.6 \text{ bits}$
---	--

The uncertainty falls off rapidly with knowledge of preceding context, an effect basically due to the finite size of the corpus but important for the processing of such corpora.

B. Comparison of Indexing Vocabularies and Message Indexing Coverage

No matter how efficient the search logic, retrieval performance is of course limited by the quality of the original indexing of messages. Recognizing this, we have concerned ourselves with comparing the properties of our indexing vocabularies: the manual ones (GE-0A and

GE-1A) and the automatic one (GE-2A), and with investigating in detail how they are used to index messages. We have been concerned with the following basic questions:

- (1) How many of the GE-2A automatic indexing terms are either identical with or close cognates of terms in the manual UNITERM set used to index the parent GE-0 collection?
- (2) Focusing on those terms that were used in indexing both the GE-1 and GE-2 collections, how do the respective usage frequencies compare; i.e., did the manual- and machine-indexing procedures tend to use a given term the same number of times?
- (3) What is the overlap, on a message-by-message basis, of the GE-0S and GE-2S index sets assigned to that item?* Specifically, what is the conditional probability that a UNITERM assigned to a message will also be a GE-2A machine term present in the text? Conversely, what is the probability that a GE-2A term assigned to an item will also be assigned as a UNITERM?
- (4) Focusing more closely on how the vocabularies are actually applied to indexing individual messages, how do the manual (GE-0S) and automatic (GE-2S) index sets compare in their depth of coverage of the conceptual material contained in the messages? Does one of the indexings give better coverage, or use fewer or better terms? How do they compare in producing spurious assignments; i.e., indexings of concepts not in fact contained in a message?

1. Study of Inclusion of GE-2A Terms in the Original GE-0A UNITERM Set

The GE-0A index set (4,824 terms) is almost five times the size of the GE-2A vocabulary. It seemed sensible to begin by asking as a first question, whether the one set includes the other. That is, how many of the GE-2A terms are either identical with or close cognates of terms in the UNITERM set of the parent GE-0 collection? To answer this, we classified the 999 GE-2A terms into the following four categories:

* We introduce the diacritical "S" to designate a surrogate, i.e., an index set assigned to a particular item.

- C The GE-2A term is either a UNITERM or is the plural of a UNITERM.
- X-C The GE-2A term is not in the UNITERM list but has the same morphological stem as a UNITERM and is closely related to it in meaning. (Examples: accelerate and acceleration; grow and growth; capable and capability.) Such terms will be referred to here as morphological cognates.
- X The GE-2A term is neither in the UNITERM list nor a morphological cognate of a UNITERM.
- X-C Not otherwise classified. There are 39 unusually short terms in the GE-2A list which were not classified because of the difficulty of determining their meaning out of context. These terms contain between one and three letters, and in many cases are abbreviations of words in the UNITERM vocabulary. Percentage figures given here exclude these 39 terms.

We determined both the fraction of the GE-2A vocabulary and the number of token usages in text accounted for by words in the first three categories. The results are displayed in Table V-4.

We found it of interest that such a high percentage ($730/999 = 73\%$) of the GE-2A vocabulary is formally included (Category C) in the UNITERM vocabulary. Moreover, these terms are found to account for about 85% of the actual token usages (postings) in the automatic indexing. This high degree of inclusion suggests that a simple machine-derived indexing vocabulary need not be very different in formal makeup than a UNITERM vocabulary consciously selected by indexers. It is of interest to analyze further the 15% (X or X-C) terms which are not in the UNITERM vocabulary.

Category	Number of GE-2A Terms in This Category	% of Reduced GE-2A Term Set	Aggregate Total Usages In GE-2 Text	% of 207,508 Usages
C	730	76%	176,218	85%
X-C	165	17%	22,332	11%
X	65	7%	8,958	4%
			<hr/> 207,508	

GE-2A Terms Included in the UNITERM Vocabulary

TABLE V-4

a. Morphological Cognates

Most of the UNITERMS are nouns, but many of the morphological cognate X-C terms in the GE-2A vocabulary are not. We broke the list of X-C cognate words down according to whether the GE-2A term is likely to be primarily a noun, a verb, a verbal (i.e., a participle) or a modifier. The figures for breakdowns into these categories are:

X-C Sub-Category	Number of GE-2A Terms	% of GE-2A Vocabulary	Aggregate Total Usages In GE-2 Text	% of 207,508 Usages
Noun	31	3%	3,770	1.8%
Verb	9	1%	1,202	0.6%
Verbal	93	10%	12,565	6.1%
Modifier	32	3%	4,795	2.3%
Total X-C	<u>165</u>	<u>17%</u>	<u>22,332</u>	<u>10.8%</u>

Morphological Cognates

TABLE V-5

We have compared the 165 X-C words with their UNITERM cognates. In the majority of cases, the differences between the text term and UNITERM is primarily one of grammatical form rather than meaning. For the large majority of X-C word forms, its morphological cognate in the UNITERM list can, in our judgment, be construed as an acceptable reduction of the X-C word to canonical form.

b. Test Words Without UNITERM Cognates

The 65 GE-2A terms which are neither UNITERMS nor have UNITERM morphological cognates can be broken down according to whether they are noun, modifier, verb, verbal, or adverb, giving the following data:

Sub-Category of X	Number of GE-2A Terms	% of GE-2A Vocabulary	Total Usages	% of 207,508 Usages
Noun	19	1.9%	1,922	0.9%
Verb	9	0.9%	1,748	0.8%
Verbal	13	1.3%	2,786	1.4%
Modifier	21	2.1%	2,305	1.1%
Adverb	3	0.3%	197	0.1%
Total	<u>65</u>	<u>6.5%</u>	<u>8,958</u>	<u>4.3%</u>

GE-2A Test Words Without UNITERM Cognates

TABLE V-6

Inspection of the words which fall into this category (X) shows them to be very "general" and "vague" words that appear to be of little value when used as one-word search requests for retrieving messages from the present specialized collection; they are not particularly content-bearing -- we sometimes refer to them as "colorless" words. Possibly this is the reason why indexers of the GE collection chose not to make them UNITERMS despite their tendency to be used repeatedly in text. However, such terms are of value when combined in a search request with other terms having more specific denotations.*

2. Comparative Usage Frequencies of Terms Common to the GE-1A and GE-2A Vocabularies

The correlation of usage frequencies of terms resulting from manual and machine assignment processes yields further insight into the similarities of the resulting indexings: do the two processes tend to assign a given term equally often?

A total of 240 terms overlap the GE-1A and GE-2A vocabularies, and there are an additional 41 close nearly identical cognates (e.g., "circle" in GE-1A and "circular" in GE-2A). The correlations of usage

* This also sometimes holds when "colorless" words are combined with each other: for example, TAKE and OFF seem colorless alone but combine to form TAKE OFF which, in our collection, is used in the context of airplanes.

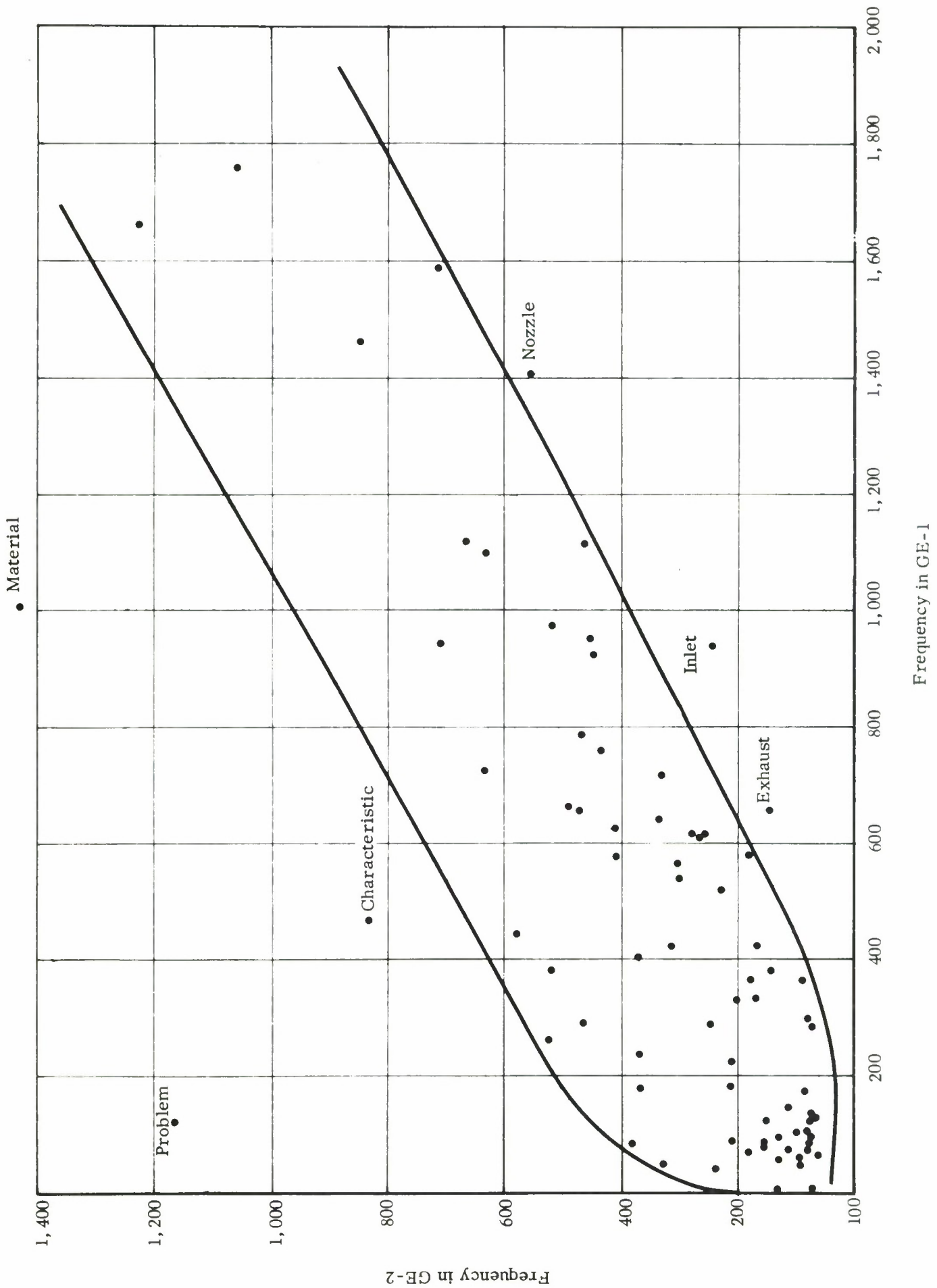


TABLE V-7 OCCURENCE FREQUENCY CORRELATION OF TERMS COMMON TO GE-1 AND GE-2 VOCABULARIES --
 Redundant low-frequency points are omitted

frequencies of a sample of the 240 overlapping terms is shown in Table V-7. Words with particularly anomalous usages are labeled: the three words with comparatively high auto-indexing frequencies are "problem," "characteristic," and "material," all rather general terms. Those with comparatively low auto-indexing frequencies are "exhaust," "inlet," and "nozzle," terms which seem to be closely related semantically.

With some scatter, the terms appear to be used with comparable frequencies in the two indexing modes.

3. Indexing Coverage and Quality

The previous discussions treat over-all properties of the indexing vocabularies, independently of how they have been applied to index individual messages. Essentially we have shown that most of the GE-2A auto-indexing vocabulary is included in the GE-0A vocabulary, and that there appears to be some degree of correlation of usage frequency of a term in the two indexings. However, we have said nothing yet as to the degree to which the manual and machine indexing procedures tend to assign the same terms to a given specific message. We felt it to be important for us to seek additional insight as to whether our automatic indexing is of comparable over-all quality with the original manual indexing, and therefore we conducted some additional studies of how individual messages are actually indexed by the various vocabularies. We first worked with a sample of only twelve specific messages from the parent GE-0 collection; later we enlarged the sample size to 33 messages and finally to 50 messages. The results, described in parts (a), (b), (c) below, remained essentially the same for all three sample sizes, and we therefore concluded that study of additional messages was unnecessary.

a. Vocabulary Usage Overlap

The first question we were concerned with was: What is the formal overlap of the GE-0S, GE-1S* and GE-2S index sets on a message-by-message basis? We compared the index sets assigned to the 50 messages in the sample and counted, for each message, the sizes of certain term sets and subsets. We thus obtained the following averages for the given sample:

* Throughout this section we report overlaps relative to the GE-1S indexing as well as the others. These data are, however, based only on the first twelve messages examined. It was clear at that point that the comparison between GE-0S and GE-2S was far more interesting, and study was extended for those vocabularies.

α = number of GE-0A terms assigned = 20.6

β = number of GE-1A terms assigned = 14

γ = number of GE-2A terms assigned = 18.2

δ = number of terms assigned by both GE-0S and GE-2S which are identical = 8.3

ϵ = number of terms assigned by both GE-1S and GE-2S which are identical = 6

From these figures we note that if a GE-2 term is assigned by machine to a message, the observed probability is $\delta/\gamma = 0.46$ that the same term was originally assigned by a human indexer as a UNITERM and, conversely, if an indexer assigned a UNITERM to a message the observed probability is $\delta/\alpha = 0.40$ that it was also assigned that term by machine. If cognates are counted, the overlap is of course higher.

b. Conceptual Coverage of Indexing

It was equally important to assess how well the different index sets assigned to a message covered the conceptual contents of that particular message. To investigate this question, we inspected each of the 50 messages individually, interpreting and comparing the text of the message against the three index sets assigned to it: GE-0S, GE-1S, and GE-2S. We prepared five lists for each message, each list corresponding to a type of indexing discrepancy and each containing names of concepts which are discrepancies of that type for the given message. A typical listing is exhibited in Table V-8.

We averaged the number of discrepancies of each type over the 50 messages investigated and obtained the following figures:

List A -- Message Concepts Missed by the GE-0 Indexing

1. Environmental Test Facility
2. Development

List B -- Message Concepts Missed by the GE-1 Indexing

1. Environmental Test Facility
2. Development
3. Test Cell Complex
4. Control Apparatus

List C -- Message Concepts Missed by the GE-2 Indexing

1. Snap
2. Environmental
3. Remote (Control)

List D -- Spurious Concepts in the GE-0 Indexing

1. Radiation*
2. Pit
3. Shield
4. Atmosphere
5. Hot
6. Television
7. Ventilation

List E -- Spurious Concepts in the GE-1 Indexing

1. Radiation*
2. Pit
3. Shield

* The majority of the "spurious concepts" are ones which were felt to be neither referred to nor implied in the text of the message. A few of the "spurious concepts" (in this case, only "radiation") were felt to be possibly suggested by the contents of the message in some way. The results described here are essentially the same, whether or not these asterisked items are included among the clearly spurious ones.

Example Worksheet for Message #59512
Names of Concepts Which were Judged to be in
Discrepancy Between Actual Textual Content of
Message and Indexing Assignment

TABLE V-8

<u>List</u>	<u>Number</u>	<u>Average Number Per Message Of:</u>
A	3.5	Significant concepts mentioned in the text of the message but not indexed in the GE-0 UNITERM indexing
B	5.5	Significant concepts mentioned in the text of the message but not indexed in the GE-1 UNITERM subset vocabulary
C	3.4	Significant concepts mentioned in the text of the message but not indexed in the GE-2 auto-indexing vocabulary
D	6.5	Significant concepts* named in the GE-0 index set but not treated in the text of the message (apparently spurious)
E	1.9	Significant concepts* named in the GE-1 index set but not treated in the text of the message
F	4.1	Significant concepts* named in the GE-0 index set but neither treated nor suggested by the content of the message (clearly spurious)

From these figures we noted that the observed probability that a concept in a message will fail to be indexed is about the same (within observational error) in either the GE-0 manual indexing or the GE-2 automatic indexing. We feel that this is rather remarkable, given the fact that the automatic indexing vocabulary contains only 999 distinct terms as contrasted with 4824 distinct terms employed in the manual indexing vocabulary. On the average, we found that about one concept per abstract is missed in common by both the GE-0 and GE-2 indexings (i.e., the average number of elements in the intersection of lists A and C). That is, the two indexings have a tendency to miss the same concepts.

Looking now at spurious concepts (list D) assigned in indexing, we note that the two indexing systems differ in a major way -- manual indexing assigned an average of 6.5

* These were "single-term" concepts. If the term "AIRPLANE" were assigned to an abstract that treats only the insides of a piston engine we would consider the term conceptually spurious. Basically this would occur when we failed to detect a plausible reason for the presence of the term in the index set for the abstract. We did not treat term-combinations at all in D, E, F, and false combinations of terms were not studied.

possibly spurious terms per message (of which 4.1 are clearly spurious), while machine indexing (of course) assigned none. The probable reasons for this are discussed in part (c) below.

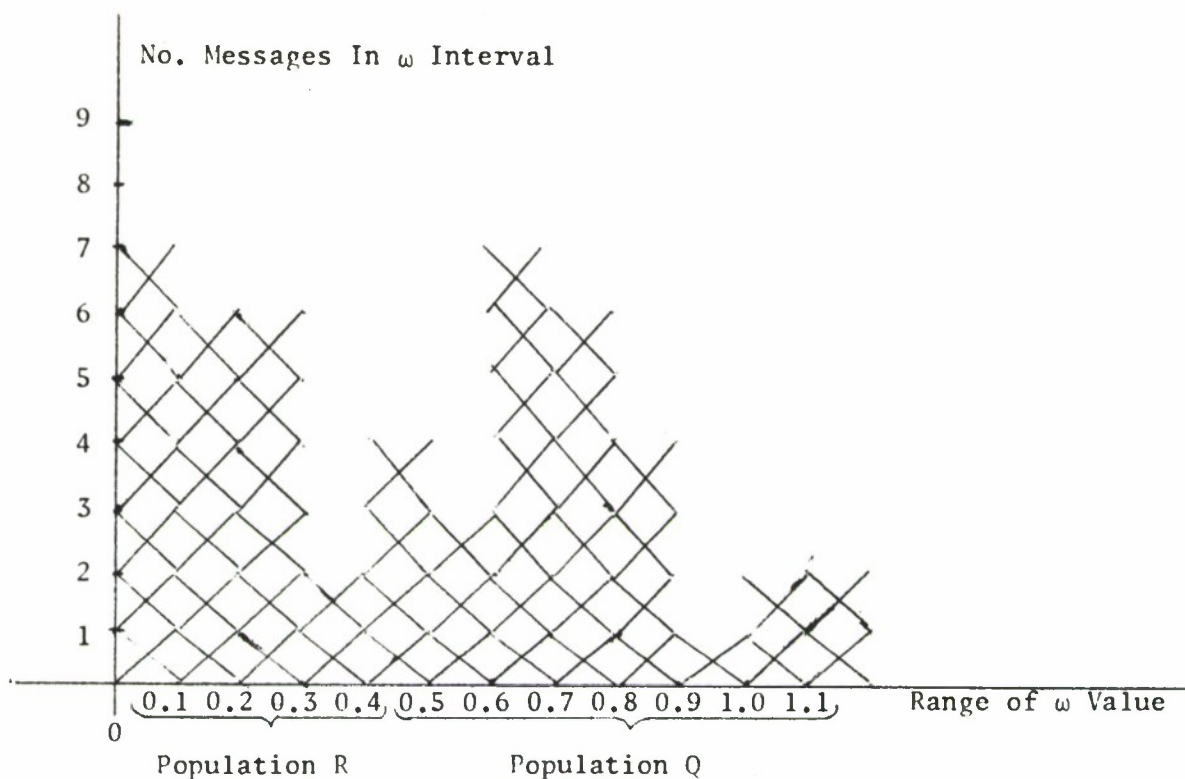
c. Analysis of Spurious Manual Term Assignments

We infer that there is a good reason for the high observed rate of spurious manual indexing assignments; namely, because the indexers were trying to represent parent documents (reports, journal articles, etc.) rather than the abstracts. In fact, the variance in the number of spurious terms was so high from message to message that we suspected we were dealing with two distinct populations of messages: one population (we call it R) in which parent documents were both indexed and abstracted at the same time by the same person, and another population (called Q) in which author or review journal abstracts were used but in which indexing was still done considering both the parent document and the abstract. We could not tell whether this suspicion was correct or where the abstracts came from by merely looking at them. However, the fact that two distinct populations can be discriminated was obvious from the bimodal nature of the distribution of spurious term set size. This can be exhibited several ways; for example, the plot of the ratio

$$\frac{\text{\# spurious manual assignments}}{\text{\# nonspurious manual assignments}} = \omega$$

against number of messages (see Table V-9A).

We arbitrarily picked a cutoff as indicated to define the two populations R and Q, and entertained the explanation already given. We reasoned that if we wished to consider manual indexing of messages -- rather than of parent documents -- we should recompute our data considering only population R, this population corresponding more to a situation where the indexer intends to index the message rather than the parent document. The results of computing the major measures for the two subpopulations R and Q as compared against the whole are exhibited in Table V-9B.



Distribution of Spurious Term Set Size

TABLE V-9A

Average Per Message Of:	Entire Sample	Subpopulation R Indexer-Abstracted*	Subpopulation Q Author-Abstracted*
α = $\frac{\text{\# GE-0 terms}}{\text{message}}$	20.6	16.8	23.3
γ = $\frac{\text{\# GE-2 terms}}{\text{message}}$	18.2	18.4	18.0
δ = $\frac{\text{\# terms assigned by GE-0 and GE-2 which are identical}}{\text{message}}$	8.3	9.7	7.3
Λ = $\frac{\text{\# message concepts missed by GE-0 indexing}}{\text{message}}$	3.5	3.4	3.5
C = $\frac{\text{\# message concepts missed by GE-2 indexing}}{\text{message}}$	3.1	3.4	2.9
D = $\frac{\text{spurious concepts in the GE-0 indexing}}{\text{message}}$	6.5	2.1	9.6
F = $\frac{\text{clearly spurious concepts in the GE-0 indexing}}{\text{message}}$	4.1	1.0	6.4
* See text.			

Comparison of Measures for Entire Sample
and Subpopulations R and Q

TABLE V-9B

To summarize the important results of this subsection:

Both original manual indexing and machine indexing appear to cover the conceptual material in individual messages with comparable thoroughness. However, in our collection, manual indexing has resulted in numerous indexing assignments that are spurious with respect to the content of the message. The machine indexing on the other hand does not produce spurious term assignments of this kind. Finally, the machine indexing achieves comparable coverage using a much smaller total number of terms (by a factor of almost 5).

It appears that the messages can be split into two subpopulations; in the first, the abstract was probably written by author or someone other than indexer; in the second, the abstract and index set were probably prepared simultaneously. The second subpopulation exhibits a reduced rate of spurious manual indexing assignments, but otherwise the statements above continue to hold for both subpopulations.

4. Extrapolations of Overlap Results and Estimates of Collection Parameters

We found it desirable to extrapolate the results of our various overlap studies in the interests of obtaining certain numbers that describe over-all properties of our experimental collections. We devote Appendix C to summarizing our best estimates for the overlap parameters of the vocabularies, the message items, and the various index sets.

Before embarking upon Section VI and our experiments in comparative evaluation, we complete our description of the experimental situation with a brief description of the retrieval tools available for our use.

C. Retrieval Searching Tools

A number of different search procedures are available to us, but all depend on use of certain searching tools which have been prepared using machine methods. For each of the two collections, tools which are available consist of certain machine-printable lists, magnetic tapes, and computer searching programs. The printed lists are:

- (1) An alphabetical list of terms, with frequency of occurrence; this list is also available in order of decreasing use frequency of the terms.

- (2) A printed list of the terms in alphabetical order, with each term followed by a list of the numbers of the messages indexed by that term (has been actually printed for GE-1 only).
- (3) A printed list of message numbers in accession order, with each message number followed by a printed list of the terms which index that message (has been actually printed for GE-1 only).
- (4) A printed list of the actual messages in order of accession numbers (can be printed for either corpus but has not been yet because of excessive length).
- (5) A printed list of term associations. Main headings are terms in alphabetical order, and under each term are listed up to the first 50 associated terms, in decreasing order of association. The association matrix we have experimented with for corpus GE-1 is $I + K + K^2$, and that for corpus GE-2 is $\sqrt{A} KA$. Initial portions of four typical lists for the GE-2 collection are exhibited in Table V-10.

ANALOG	ARC	BENDING	BORON
computer digital simulation amplifier computing nonlinear accuracy circuit automatic servo loop exchanger electrical transient error computation	electrode case welding tungsten melting weld plasma cracking molybdenum furnace intensity electric sec metallograph welded casting	torsion moment loaded curvature shell rectangular column beam buckling deflection transverse combined loading supported rod theorem	silicon carbide acid cobalt compound beryllium lithium zircononium carbon polymer oxide graphite magnesium cermet spray decomposition

Initial Portions of Four Typical Association Lists
 $\sqrt{A} KA$ Matrix, GE-2 Vocabulary

TABLE V-10

Only the first of these lists is small enough to be used conveniently as a search tool; the others occupy large and heavy books, and are useful only for occasional references. For each collection, the following magnetic tapes are available:

- (1) A "dictionary" tape, used to convert term names into term identification numbers and vice versa.
- (2) A "C-matrix" tape which lists, for each message number, the identification numbers of the terms used to index that message.
- (3) An "association" matrix tape which is a machine-readable version of the term association printed list. While we ordinarily print only the 50 top valued association to any term, the tape entries usually contain many more than 50 associates to any word.
- (4) A tape containing the texts of the messages in serial number order.

The two computer programs we principally use in conducting retrieval are designed for efficient operation on an IBM 1401 computer (minimal configuration: 4 tapes and 8K of core memory), and have provisions for batching multiple inquiries. The first program (known as "Phase I") retrieves word association profiles; the second ("Phase II") retrieves actual messages.

The Phase I program accepts up to several dozen queries at one time; each is regarded to be a separate batch and results in separate outputs, although all are processed at once. A query can consist of full text in unconstrained natural language form or, alternatively, only of selected terms. A query is shown in machine input format in Table V-11A. The requestor has the option of assigning each input term a positive or negative weight; otherwise equal weights are automatically assigned to all input terms. The program uses only query terms present in the dictionary tape for the collection concerned. Query terms not found in the dictionary are listed and ignored in subsequent processing. The Phase I program searches the association matrix tape and prints out, in decreasing order of association strengths, an over-all word association profile for each of the given queries, using the linear associative retrieval algorithm.* Also, there is an option whereby the program prints out individual word association profiles for each separate word in a query. A separate punched-card deck is generated for each association list, a single card per term. A typical run of the Phase I program may be for 15 separate input inquiries and produce 15 corresponding inquiry association lists and punched-card decks. This would

* That is, the machine prints out the words and corresponding values of $R = KQ$, where Q is the vector of query words, K is the association matrix used and R is the vector of values of associated words.

Query Phase I

THRESHOLD .003000

ASSIGNED BATCH NUMBER 6

<u>Query Terms</u>	<u>Input Weight</u>	<u>Normalized</u>
COSMIC RADIATION	1000	3125
METAL	1000	3125
RADIATION	1000	3125
ATMOSPHERE	1000	3125
SYSTEM	1000	3125
PRESSURE	1000	3125
TEMPERATURE	1000	3125
INTENSITY	1000	3125
ENERGY	1000	3125
GAS	1000	3125

Typical Input Query to the Phase I Program*

TABLE V-11A

probably require about 45 minutes of time on our 1401, about three minutes per inquiry on the average.

Table V-11B shows the association list for the query of Table V-11A. Words present in the input query are automatically marked with an asterisk. Note that several words in the retrieved association list derive high values purely due to association (e.g., "meteorology," "vapor," "counter," "stratosphere," etc.).

* This particular example of the program's input as well as the association output shown in the next table are based on a small sample of the NASA collection. The terms are all from the NASA index term vocabulary, which contains word strings as terms as well as single words. The program does have provision for variable input weights despite the illustration (including negative weights) and in the output listing in Table V-11B, some of the + signs could be - when a negative input weight is used.

Computed Weight	Word	Computed Weight	Word
0.1187+	METEOROLOGY	0.0562+	ARGON
0.1125+	VAPOR	0.0562+	IONIZATION
0.1093+	INTENSITY*	0.0531+	COMPOSITION
0.1062+	ATMOSPHERE*	0.0531+	DIFFUSION
0.1000+	COSMIC RADIATION*	0.0531+	HIGH ENERGY
0.0843+	COUNTER	0.0531+	HIGH TEMPERATURE
0.0843+	STRATOSPHERE	0.0500+	PARTICLE
0.0843+	TEMPERATURE*	0.0468+	DETECTOR
0.0843+	VARIATION	0.0468+	FLUX
0.0781+	COSMIC	0.0468+	OXYGEN
0.0781+	TELESCOPE*	0.0437+	BOILING
0.0750+	RADIATION*	0.0437+	COEFFICIENT
0.0718+	GAS*	0.0437+	CONDENSATION
0.0687+	COMPONENT	0.0437+	COUPLING
0.0687+	ENERGY*	0.0437+	SYSTEM*
0.0687+	PRESSURE*	0.0437+	THICKNESS
0.0687+	NEUTRON	0.0437+	SOLID
0.0656+	UPPER	0.0437+	SPECTRUM
0.0656+	PROTON	0.0406+	ABSORPTION
0.0656+	RADIO	0.0406+	FUEL CELL
0.0656+	NITROGEN	0.0406+	FUEL ELEMENT
0.0625+	AIR	0.0406+	PHYSICS
0.0625+	CESIUM	0.0406+	PROPERTY
0.0625+	SOLAR	0.0406+	REACTOR

* Terms in the input query are asterisked by the program.

Association Profile Produced by the Phase I Program
For the Input Query of Table V-11A

TABLE V-11B

The Phase II program resembles the Phase I program, only it retrieves messages rather than words; it can also process several queries simultaneously. The input query consists of a set of terms on punched cards with assigned weights. The terms could be original request terms and weights might be assigned to them manually, but typically this input is the association profile resulting from a Phase I run (possibly modified to reflect human guidance of the search). The program generates a weight for each message which is basically a normalized sum of the input weights of all the terms it shares with the input term list. At the completion of the run, a listing of the retrieved messages is printed out for each inquiry; each listing is presented in

decreasing order of the computed message weights. Messages are listed in the format of Table V-1.

The retrieval options mentioned in Section I and discussed further in Section VI work as follows:

- (1) Fully Automatic Associative: The full text of the query is input to the Phase I program. All words are given equal initial input weights. The topmost portion (usually the first 100 terms) of the output deck of the Phase I program is fed directly to the Phase II program -- complete with machine-computed weights.
- (2) Selected Associative: The output deck of the Phase I program is interpreted and inspected, and only cards corresponding to associated words which seem to be pertinent to the query are retained as input to the Phase II program. Phase I machine-computed weights are retained and used in the Phase II input.
- (3) Reweighted Associative: The output deck of the Phase I program is interpreted and inspected, and certain cards containing associated words are selected either because they correspond to things the query is about or because they clearly relate to things the query is not about. These words are then reweighted, according to the human requestor's estimates of the positive or negative value a term has with respect to the query. The new weights are used in the Phase II input list.
- (4) Conventional Coordinate: The query is fed directly into the Phase II program with all input words given equal weights; no associations are generated or used.
- (5) Frequency Weighted Coordinate: The query is fed directly into the Phase II program; however, all words are previously weighted (using an automatic process) with weights proportional to the reciprocals of their frequencies; no associations are generated or used.

A system of auxiliary computer programs is used for various steps of data preparation, testing and evaluation analysis and, in addition, a comprehensive system of programs exists for automatic indexing and other steps of preparation of natural language message bases for automatic associative searching.*

One particular program of interest performs retrieval of content-bearing word strings; we call this our Phase III program. It

* See Technical Note CACL-13, (26).

Query Content: Adhesive bonding of fiber glass to metal and shear testing of the resulting joints.

Samples of Retrieved Content-Bearing Units: Phase III Program:

fiber glass	shear stresses
glass fiber	shear strain
glass laminate	stress shear
	stress rupture
adhesive bonding	elastic shear
adhesive bonded	elastic stresses
structural adhesive	shear testing
resin bonded	strain tests
structural sandwich	yield stress
structural joint	shear deformation
adhesive strength	shear strength
resin laminates	rupture stresses
metal adhesives	creep stress
metal joints	shear modulus
metal bonding	tensile testing
structural laminates	shear buckling

Examples of Content-Bearing Pairs
Retrieved by the Phase III Program
(Approximately 400 additional pairs beyond those
listed here were retrieved for the given query,
all with apparently comparable relevance.)

TABLE V-12

is similar to the Phase II in that its input consists of an association profile. Its output consists of retrieved content-bearing word strings (CBUs) which were originally obtained from the corpus through statistical extraction procedures.* We have not yet used the output of this program for systematic evaluation tests. Nonetheless, it appears to us to be potentially very useful for identification of alternative formulations for requests. Some examples of pairs retrieved using the Phase III program are exhibited in Table V-12.

* These procedures are treated in detail in Technical Note CACL-18, (2).

SECTION VI EXPERIMENTAL RESULTS

Learning About Automatic Message Retrieval

This section treats some of our experimental investigations on the GE-2 collection and assesses what we have learned about automatic message retrieval so far. The studies described are mainly concerned with the observable properties of queries, the relationship between these properties and the choice of the appropriate search options that might best be employed, and with expected search performance under various query conditions. While we have had to be selective in picking the studies to be discussed, we have accumulated much additional evidence that does not lend itself so easily to formal organization. This additional experience corroborates our main findings, and additional details about several of the more important tests can be found in the Technical Notes as indicated in the text.

We have organized the section into four subsections. Subsection A relates to the processing of subject-heading type queries, and Subsection B to the processing of long, textual queries. Subsections C and D are briefer and relate respectively to experiments involving the use of multiple evaluators and to the extrapolation of our results to manually-indexed collections. General considerations -- e.g., the circumstances under which subject-heading and full-text queries may be useful -- have already been discussed in Section III.

A. Retrieving on Subject-Heading Type Queries

We begin our discussion of experimental searches using short subject-heading queries by reviewing some results on automatic methods for the discovery of content-bearing word strings in the texts of messages; such word strings are possible subject-heading terms. We pass then to a description of our first investigation, which focuses on a large population of subject-heading queries drawn from an operational retrieval environment (NASA's). These are regarded as illustrative of the queries that might be posed to the experimental prototype in practice. We consider the questions: How many of these queries can be handled by our GE-2 experimental system, and how? What are the distributions of some of the statistical parameters of these queries (pair frequency and "cohesion" are defined below)? We next choose some subject-heading queries with differing parameter values, perform coordinate retrieval search with them, evaluate the retrieved messages for relevance, and investigate the relationship between pair frequency, cohesion, and expected precision of search. Having learned about the relationship between the observable query parameters and the quality of the resulting search, we turn to an investigation of the word association profiles retrievable as a result of searching with subject-heading type queries. We pay particular

attention to the number of pertinent associated terms, again as a function of query parameters. Finally, the performance characteristics curves of subject-heading queries using coordinate and associative search options are developed, interpreted, and compared.

1. Background on Content-Bearing Units (CBUs)

When we started the retrieval experiments discussed here, we had for some time been using the body of text drawn from the GE-2 collection for purposes of word string analysis experiments. These investigations have been conducted under a related line of investigation whose objectives and those of our evaluation work are mutually supplementary. Under the other investigation, we have been concerned in part with the development of practical methods for recognition of statistical-semantic structures present in natural language corpora and exploitation of these structures for retrieval.

The most comprehensive available description of our word string analysis investigations appears in our 93-page Interim Report, Towards the Use of Natural Language Structure in Automatic Message Retrieval.^{*} The emphasis of that report was on:

"...the testing and validation of machine techniques for discovering high-precision indexing units for use in associative retrieval. It documents our findings that relatively simple procedures are feasible for the fully automatic identification of concept-denoting word strings..."

For the present purposes it is useful to review the principal concepts in that work. Two statistical parameters of word strings have great importance; they are "pair frequency" and "cohesion." For two-word strings, pair frequency f_{ab} is simply the number of occurrences of the word string ab in the corpus (the text of the entire message collection). Cohesion is defined as

$$C_{ab} = \frac{f_{ab}}{f_a \cdot f_b} N$$

where f_a and f_b are the occurrence frequencies of the individual words a and b , and N is the number of word tokens in the corpus. In our experiments^{**} for discovering useful strings, pair frequency and cohesion were used as criteria for classifying the strings encountered

^{*} Technical Note CACL-18, (2).

^{**} Section IV of the Interim Report (2).

in the text as to whether they are (or are not) content-bearing units (CBUs). A number of different tests were done using different data samples, different classification criteria, with comparative judgments being made by human subjects. These tests supported the following observations:

- (1) The notion that certain word pairs selected from the text are concept-bearing units is meaningful, since humans can display a reasonably high degree of consistency in so classifying such pairs.
- (2) Certain very simple procedures are quite effective for selecting concept-bearing word pairs from text. For example, choosing pairs not containing exception "function words" such that $f_{ab} \geq 3$ and $C_{ab} \geq 20$, we found:
 - (a) For the 10,000-message GE-2 corpus, roughly 3000 strings are selected.
 - (b) The probability that a selected string is indeed a content-bearing unit is about 0.93.
 - (c) About 19% of the machine-selected strings are also subject headings used in the NASA Scientific and Technical Aerospace Reports (STAR) periodical and associated document retrieval system. In contrast, less than 1% of the strings present in the text (before applying the selection procedure) are subject headings of this kind.
 - (d) Making selection criteria more stringent decreases the number of strings selected but increases the percentage which are truly content-bearing as well as the percentage which are STAR subject headings; i.e., requiring $f_{ab} \geq 6$ as well as $C_{ab} \geq 20$, virtually all selected strings are concept-bearing and 34.7% of them are NASA headings.
 - (e) Almost all of the selected content-bearing units appear to be just as suitable to be descriptive subject headings as those used in STAR.

These results suggested that knowledge of such text statistics as pair frequency or cohesion could, under certain conditions, provide an objective basis for estimating or evaluating the performance of a retrieval system. Such an approach is pursued in the present chapter, and the key questions are:

- (a) Given a population of subject-heading queries (as exemplified by the roughly 9,000 two-word subject headings used in STAR) for those which out GE-2A vocabulary can handle, what are the distributions of pair frequency and cohesion in our collection?
- (b) What is the relationship between the pair frequency and cohesion of a subject-heading query and the expected performance of the resulting search, using simple coordinate retrieval logic?
- (c) How can these relationships be used to help in selection of a best search strategy, given a certain specific subject-heading query?
- (d) What are the consequences of decomposing a long, full-text query into its subject heading components prior to searching?
- (e) Is the retrieval of association profiles that contain subject headings likely to be a useful intermediate step in the retrieval process under certain conditions?

2. STAR Subject Headings as Queries

As discussed in Section III, we wished to evaluate the performance of our system against realistic queries resembling those which are known to be useful within an operational retrieval context. We reasoned that the NASA documentation system would be an excellent source for a population of such queries because, like our GE collection, it treats the general area of aerospace technology. Of course, there are not only similarities but also important differences between the NASA collection and our experimental GE-2 collection; namely, (a) the NASA collection is focused on certain specific areas like outer space, rocket technology, and space medicine, of concern to NASA, while the GE-2 collection is focused on other specific areas of concern to the Division of General Electric responsible for compilation of the parent GE-0 collection; namely, the design and testing of air-breathing jet engines, (b) the NASA collection is up to date, while the experimental GE-2 collection cuts off in 1962, and (c) the NASA collection is much larger than even the parent GE collection and goes into many matters in much more detail.

There are over 10,000 multiple-word subject headings employed regularly in the STAR publication, and these headings are also usable as machine search terms in the computerized documentation system operated by NASA. Of the total, roughly 8,800 are two-word headings, and we focused on these, primarily because the text statistics relevant to retrieval are simplest for the two-word strings. We are confident that

our observations described below continue to be valid for three-word and longer subject headings.

a. Method of Analysis

We have available to us a listing of all NASA machine searching terms ordered according to their NASA use frequency; i.e., in the decreasing ordering of the number of NASA documents posted to them. We regarded the list to be divided into ten frequency intervals, and we randomly sampled a set of two-word headings from each interval.* Sampling intensity was variable, being greatest (exhaustive) for the interval with terms having highest usage frequency (i.e., for the 122 headings used by NASA to index 197 or more documents each).

Because of the importance of the findings developed from the study of this sample, we felt it important to investigate the reliability of our sampling procedure in some depth. This analysis is included in Appendix D.

Each heading sampled was checked against vocabulary lists for the GE-2 collection. For those headings having both words in the GE-2A vocabulary, pair frequency f_{ab} and cohesion C_{ab} were then obtained from other printed lists. A few of the smaller samples which appeared to yield anomalous values were enlarged. Altogether, 410 headings were sampled and studied, and the results of this study were extrapolated to the entire set of STAR queries to yield the results described below.

b. Main Results

Extrapolation from the samples provides the following summary estimates for STAR two-word subject headings:

<u>Event</u>	<u>Probabilities</u>	
	<u>By Query Usage</u>	<u>By Query Type</u>
(a) At least one query word in GE-2A	0.93	0.88
(b) Second query word is in GE-2A	0.80	0.73
(c) Both query words in GE-2A	0.55	0.29
(d) Neither query word in GE-2A	0.07	0.12

* Details are given in Technical Note CACL-31, (27).

In Appendix D we show that these probabilities are within sufficiently small probable lower bounds to justify the findings which follow. The rightmost column shows probabilities based on query types, not usage. That is, if one selects a two-word subject heading at random from a list of all STAR subject headings,* the estimated probability that at least one query word is in GE-2A is 0.88, that at least the second word is in GE-2A is 0.73, and that both words are in GE-2A is 0.29. The leftmost column shows estimated query usage probabilities and takes into account the prospect that some of the subject headings (e.g., rocket engine, differential equation) might be much more likely to appear as queries than others (e.g., Vintis Theory, Cepheus Constellation). To obtain the figures, we assumed that such a propensity would be roughly proportional to the NASA posting frequency of the heading; i.e., if heading A has ten times more NASA documents posted to it than does heading B, then the probability of A being used as a subject-heading query is ten times that of B.

With respect to the adequacy of our indexing vocabulary to handle STAR subject-heading queries, our estimates lead us to expect that 7% of the time there is no choice but to completely reformulate the query because neither word is in GE-2A. For the 38% when only one word or the other is in GE-2A, some help can obviously be expected from displaying the association profile of the one word that is present. Hopefully, the profile listing might contain a suitable near-synonym of the missing word that could be used in its place. It is interesting that the first (leftmost) word can be expected to be missing 25% of the time, but the second word is missing only 13% of the time. Since the missing leftmost word in a subject-heading pair is usually a modifier, the nature of the demand that would be placed upon an association profile may well be greater for finding alternative modifiers than for finding nouns under the "one word" circumstances just discussed.**

The subpopulation of the STAR subject headings for which both words are in GE-2A was of greatest initial interest. Accordingly, we studied the distribution of the f_{ab} pair frequency and C_{ab} cohesion statistics in the various samples for this subpopulation. Some summary estimates which will be of relevance to following discussions are:

* Available, for example, in Guide to Subject Indexes for STAR, (28).

** Since modifying is a contiguity relation, this suggests that odd powers in the association matrix series should be emphasized; see ref.(1).

<u>Event</u>	<u>Probabilities</u>	
	<u>By Query Usage</u>	<u>By Query Type</u>
Both a and b in GE-2A <u>and</u>		
(a) The pair string occurs twice or more in our message collection; i.e., $f_{ab} \geq 2$ in GE-2	0.78	0.45
(b) The heading is a "System CBU"; i.e., $f_{ab} \geq 3$ and $C_{ab} \geq 20$	0.61	0.31
(c) $f_{ab} \geq 1$	0.22	0.55

These probabilities will later be related to search performance in various ways.

Further data relating to distributions of the f_{ab} and C_{ab} statistics as a function of STAR cumulative posting frequency are exhibited in the supplement to TN CACL-31. Of interest are the data for the lowest frequency STAR headings; i.e., those which have only one document posted to them in the NASA subcollection for which we have posting-frequency information.

For two-word subject headings occurring only once in the NASA collection:

<u>Event</u>	<u>Probability</u>
(a) At least one word is in GE-2A	0.92
(b) Second word is in GE-2A	0.75
(c) Both words are in GE-2A	0.17
(d) Neither word is in GE-2A	0.08
(e) Percent of those in (c) for which $f_{ab} \geq 1$	0

Recalling that the GE-2A vocabulary contains only the 999 most frequent nonfunction words in a different collection from NASA's, the consistently high overlap is noteworthy. As one would expect, brand-new subject headings in the NASA collection (by definition a new term enters when the first document is posted to it) are being formed from old and frequent words. We note that the figures presented in items

(a)-(d) are not drastically different for these new terms than for those already well-used in the collection.

3. Coordinate Retrieval Using Subject-Heading Queries

a. Issues Investigated

The next issue we investigated is a fundamental one to the whole notion of coordinate retrieval; namely, the relationship between the incidence of the component words of a query (in this case, a subject heading) within a message and the expected relevance of the message to the concept expressed by the query. We wanted to relate the expected precision and recall of coordinate searching to the pair frequency and cohesion of the subject headings in our collection. Specifically, we wanted to see if we could use cohesion and/or pair frequency to predict the coordinate search performance of a subject-heading query which is a System CBU.

b. Method of Analysis

An experiment was designed to investigate the precision resulting from doing coordinate retrieval searching using component words of subject headings as search terms and to relate this precision to the corpus frequency f_{ab} and cohesion values of subject-heading strings. Twenty-three strings were selected for use in the experiment, with the following selection criteria in mind.

- (a) All component words in the headings were in the GE-2A vocabulary.
- (b) The headings are all very clearly "concept-bearing." (Examples were differential equation, control surface, and ablation cooling.)
- (c) The heading covered a wide range of pair occurrence frequencies f_{ab} in our corpus (f_{ab} between 0 and 102), and a corresponding wide range of cohesion C_{ab} values.
- (d) Most of the headings (15 of them) are used in STAR.
- (e) Most of the headings (17 of them) are System CBUs; i.e., $f_{ab} > 3$, $C_{ab} > 20$. To economize on computer searching, several of the headings

were selected to share words in common.
All of the CBUs but one involved only two
component words.*

Coordinate retrieval requests were formulated and run in order to retrieve those messages in the GF-2 collection containing both component words of each heading in the list. For each of the 23 queries, a sample of messages containing both words in the heading was studied. Since the matter of importance under study was precision of search (percentage of messages retrieved which are relevant) it was not necessary to identify and evaluate all messages containing each pair. Altogether, 309 messages were inspected and evaluated for relevance, an average of about 13 messages per heading. Judgments of relevance were made on a conceptual rather than on a formal level. Before evaluating the messages in the sample retrieved for a given subject heading, an attempt was made to decide what constitutes relevance of a message to the concept represented by the heading. For the first two samples run, those for "low aspect ratio" and "differential equation," the criteria were set down explicitly on paper; for the latter samples they were formulated mentally. For example:

A message is relevant to the concept represented by "low aspect ratio" either

- (1) if it is about (deals directly with in a significant sense) ratios (mathematical quantities) which represent aspects (i.e., wing length/wing width) and which furthermore are low in value, or
- (2) if the adjectival concept represented by "low aspect ratio" is significant in delineating what the message is about.

Great care was exercised in evaluating the first few samples, with five distinct levels of judgment of relevance being used. It was then decided that only two levels of relevance were adequate for subsequent judgments. (This last decision applies only to the particular test being described in this subsection, where relevance is being considered with respect to a single, rather simple concept.) A coding scheme was adopted whereby a message was coded:

* TN CACL-17, (29) lists the CBU strings employed and their statistical properties, discusses the sampling procedure, and exhibits the observed data.

C: the message contains the given subject-heading word string,

or

N: the message contains members of the heading but not the heading itself as a word string,

and

R or U: the message is Relevant or Unrelevant to the concept described by the CRU.

In almost all cases when the message actually contained the heading string, it was decided that the message was relevant to the heading and it was marked CR. An example of a message which would be classified NR for "aspect ratio" would be one that read in part "...important stability-determining variables in wing design are considered with emphasis on aspect. Values of this ratio between 1 and 8 are considered..." A message containing "aspect" and "ratio" only in the following string would be classified NU: "...an important aspect of which is turbulence control. The pressure-velocity ratio in turn might..." presuming, of course, that the concept is not dealt with elsewhere in the message. In many cases a message turns out to be relevant to an explicit subject-heading request when the individual words of the heading appear widely separated in that message.

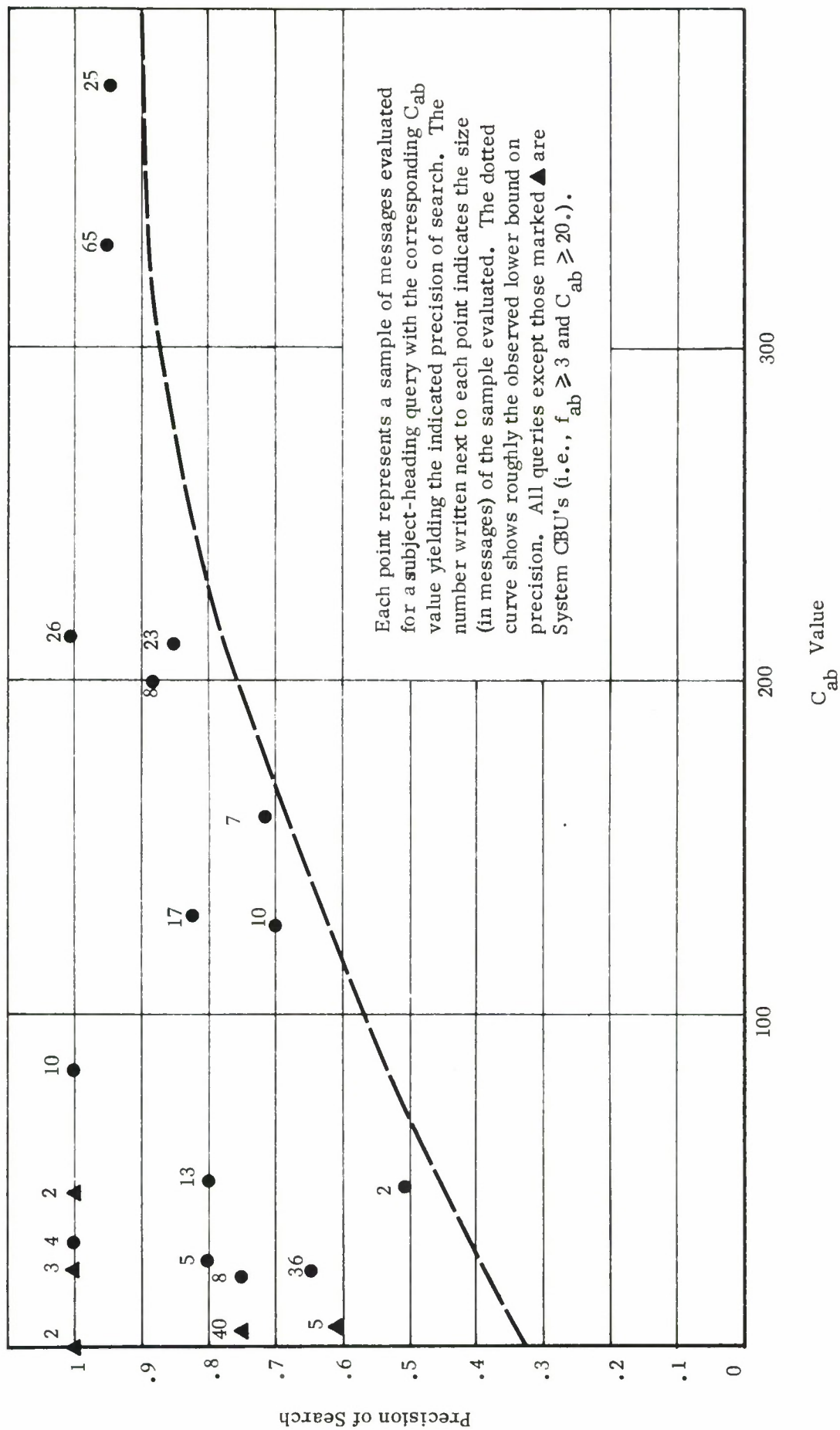
c. Main Results

After evaluating each of the 23 samples, counts were made of the numbers of messages evaluated CR, CU, NR, and NU, and two precision ratios computed:

$$P_1 = \text{Over-all Precision} = \frac{CR + NR}{CR + NR + CU + NU}$$

$$P_2 = \frac{\text{Precision for Messages Containing The Heading Words but not the Heading String Itself}}{\text{The Heading Words but not the Heading String Itself}} = \frac{NR}{NR + NU}$$

Results are exhibited in Table 1 of TN CACL-17; averages over the 23 queries studied are $\bar{P}_1 = 0.84$; $\bar{P}_2 = 0.62$. Finally, we plotted the precision of each of the 23 searches as a function of the cohesion C_{ab} value of the query subject heading and also against its f_{ab} frequency. The first such plot is exhibited in Table VI-1.

TABLE VI-1 PRECISION OF SEARCH PLOTTED AGAINST C_{ab} VALUE FOR SEARCHES BASED ON CBU COMPONENTS

The observed regularities in the data appear to support the following hypotheses:

- (1) A Two-term coordinate retrieval search employing (i.e., coordinating) the two words in a concept-bearing subject heading which is a System CBU appears to produce extraordinarily* high search precision, even for strings with low C_{ab} values. Average value over all searches is 0.84, with all values being above 0.5.
- (2) A high cohesion C_{ab} value of a subject heading implies high precision of a search employing that heading as a query. Note that for $C_{ab} > 100$, precision ≥ 0.7 ; for $C_{ab} > 200$, precision ≥ 0.85 . C_{ab} value seems to be correlated with a lower bound on precision, as suggested by the dotted line on Table VI-1; however, many subject headings which yield high precision searches may have low C_{ab} values.
- (3) On the average, 62% of the messages containing the individual words in a subject-heading string which is a System CBU but not containing the string itself are still nonetheless relevant to the concept represented by the heading.
- (4) Precision can be raised to nearly 100% by considering only messages actually containing the subject-heading string, but this entails a loss of roughly 31% of the relevant messages otherwise retrievable.**

The evidence appears unmistakably to indicate that subject-heading queries which are System CBUs yield exceptionally high precision performance for coordinate retrieval searches, particularly when the CBU has a large value of the cohesion statistic.

* The observed figure is high by comparison with the results of tests performed on actual working indexes. See, for example, p. 88 of (22) where Cleverdon et al state that "...it appears that document retrieval systems in the field of science and engineering are generally working in the range of 70-90% recall and 8-20% relevance." The 0.84 average precision observed for System CBU searches is three or four times as high as the 0.20 or 0.25 precision that Cleverdon et al consider to be the expected operating point in a typical engineering collection.

** Cleverdon et al observe (22) that a 1% improvement in precision will roughly cost a 3% drop in recall.

d. Recapitulation

Combining the results just described and those of Subsection 2 above, it appears that about 60% of the NASA STAR subject-heading queries (percentage of queries with both words on the GE-2 vocabulary, based on expected query usage) will be System CBUs and these will tend to yield relatively high precision coordinate retrieval searches, with expected precision in the order of 0.84. Moreover, since System CBUs occur at least three times in the GE-2 text, the retrieval of at least a few relevant messages is insured for such a query. The evidence indicates that both the lower bound on precision and expected precision drops with lower values of f_{ab} and C_{ab} , as does the expected number of relevant items to be retrieved. For the 40% of subject headings which are not System CBUs, neither high precision nor high recall can be guaranteed for coordinate searches on the constituent words. Finally, we observed that the queries resulting in best coordinate search precision performance (i.e., the System CBUs) are all among the most commonly used NASA STAR subject headings.

4. Association Profiles for Subject-Heading Queries

As discussed in Section III, there are two types of retrieval situations in which the display of machine-computed term association profiles might be a useful step, given a subject-heading query. These are: (a) when the query is one for which the expected performance of a coordinate search is poor, or (b) when the query is merely an initial representation of what the requestor wants and is used as an entry point with the intent of eliciting responses from the system that will aid in formulating a more appropriate query statement. For our purposes, subcase (a) subsumes the situation when one or more query words are missing from the GE-2 vocabulary as well as the situation when the subject-heading query is not a System CBU. On the whole, for about 30% of all subject-heading queries (again, based on the STAR usage statistics), high performance of coordinate searching can be predicted; in the remaining 70% of the cases, a possible recourse is the use of machine-computed associations.

a. Issues Investigated

The study to be described here was concerned with getting additional understanding of the nature of subject-heading requests, and a feeling for the nature of the word association profiles produced by them. Specifically, we were concerned with comparing the association profiles produced by requests consisting of (a) selected subject headings which are System CBUs; (b) subject headings which are not System CBUs; (c) randomly-paired GE-2 vocabulary

words; and (d) deliberately mismatched GE-2 vocabulary words. In particular, we wanted to know whether semantically cohesive subject-heading requests which are not System CBUs nonetheless tend to produce reasonable and potentially useful association profiles.

b. Method of Analysis

For purposes of conducting preliminary experiments involving associative retrieval in a real-time context, we have developed a simple time-sharing computer program for adding or multiplying together word association profiles and displaying the results.* The data base available to the program consisted of association profiles of 83 selected words from the GE-2 vocabulary, each profile truncated so that only mutual associations among the 83 words remained.

For purposes of experiment, four lists of two-word queries were generated, all queries containing only words among the 83 recognized by the time-sharing program. These lists were made up as follows:

Randomly Paired Words: Words in this list were paired together randomly, using a table of random numbers to generate the pairings (e.g., "physical airfoil").

Selected Subject Headings: Inspecting the 83 words in the selected vocabulary but without reference to any other data, the words were combined into pairs which appeared to the investigator to be reasonable subject headings (e.g., "adhesive bond").

Mismatched Vocabulary Words: Again without reference to any data other than the vocabulary list, the investigator attempted to pair words so that the words in a pair could not easily be thought of as combining to form a conceptual unit (e.g., "brittle torsion").

System CBUs: Checking against our master listings for the GE collections, pairs were identified which make up word strings which are both reasonable

* The program was run on a GE-265 computer in Phoenix, Arizona, and actuated and controled over phone lines from a time-sharing console at our office in Cambridge, Massachusetts. It is described in the Supplement to Technical Note, CACL-26, (30).

subject headings and System CBUs in our collection (e.g., $f_{ab} \geq 3$ and $C_{ab} \geq 20$).

For each of the above categories, 14 to 25 queries were developed. Each query word pair a,b was then looked up in various listings to determine (a) its backward and forward string occurrence frequencies in our collection; i.e., f_{ab} and f_{ba} , (b) its backward and forward cohesion; i.e., C_{ab} and C_{ba} , and (c) whether either ordering of the pair resulted in a NASA subject heading. The randomly paired-words were inspected to see if they defined a meaningful concept when strung together in one or the other ordering. Finally, using the time-sharing program, two different word association profiles were retrieved for 56 queries, 14 from each list; (a) "sum" profile consisting essentially of an ordered listing of all terms in the logical sum of the profiles for the individual words, and (b) a "product" profile consisting of an ordered listing of terms which are shared by the individual association profiles of both component words. Words in the product profiles were evaluated for relevance to the query. Various listing and tabulations were then prepared yielding the main results described below.

c. Main Results

The data gathered for the various strings prior to doing the profile retrievals can be summarized as follows:

	<u>No. Strings Investigated*</u>	<u>Observed Numbers</u>		
		<u>In NASA STAR</u>	<u>GE-2 System CBUs</u>	<u>With $f_{ab} \geq 2$ or $f_{ba} \geq 2$</u>
Randomly-Paired Words	50	0	0	1
Selected Subject Headings	28	5	4	6
Mismatched Words	28	0	0	0
System CBUs	50	3	25	25

More than half of the 50 strings which could be made up from the Randomly Paired Words look as if they could be

* Both forward and backward strings are counted.

perfectly legitimate subject headings (for example, "ceramic layer," "reinforced construction," etc.). What is fascinating is that despite the apparent legitimacy of these randomly-generated headings, they behave just like the Mismatched Words in the above table; i.e., they tend not to occur in either the GE-2 half-million word corpus or among the NASA STAR subject headings. Likewise, surprisingly few of the Selected Subject Headings occur in either the corpus or among the NASA headings.

To encapsulate the results so far, a random pairing of vocabulary items may well define a conceptually meaningful unit which could conceivably function as a subject heading. However, the probability that such a randomly-generated unit actually occurs as a word string in our real GE-2 corpus or among the real collection of NASA STAR subject headings is small. Even if conscious effort is devoted to pairing the words in such a manner that they make apparently reasonable subject headings (Selected Subject Headings), the chances are still about only one out of five that the heading will be in the GE-2 corpus or in the NASA STAR set of headings.

One of the first things observed while inspecting the retrieved association lists is that a word present in a "product" association list almost invariably has some degree of pertinence to the concept expressed by the subject-heading query. For example, the words in the product association list for the query "Ceramic Coating" are "ceramic," "coating," "cermet," "bonded," "adhesive," "resistant," "glass," and "metal." We did not find any word in the product profiles which could be ruled out as irrelevant to such a query. We could therefore safely assume that the number of words n_{ab} in the product profile provided a lower bound on the number of relevant associations in the normal (sum) profile for subject-heading query, since the words in the product profile appear topmost in the listing of the normal sum profile. For each of the 56 queries that were processed, we computed two measures of the effectiveness in retrieving relevant word associations, Number in the Product n_{ab} , and an Overlap Ratio, defined to be

$$r = \frac{n_{ab}}{n_a + n_b - n_{ab}} ,$$

where n_a and n_b are respectively the number of terms in the individual profiles for words a and b . If the individual profile lists are identical, then $r = 1$; if there is no overlap between them, then $r = 0$. The ratio r is a

lower bound on the probability that a word in the sum profile is pertinent to the concept expressed by the given subject heading.

The averages of these quantities for queries in the four categories are as follows:

<u>List</u>	<u>Average Number of Associations per Query</u>	<u>Average Overlap Ratio</u>
Random Pairs (all)	0.86	0.05
Random Pairs which can combine to denote concepts	1.50	0.09
Random Pairs which do not combine to denote concepts	0.00	0.00
Selected Subject Headings	3.70	0.30
Mismatched Words	0.57	0.04
System CBUs	4.36	0.28

These results were obtained from mutual association only among 83 words; i.e., we dealt with an 83 x 83 submatrix of our 1000 x 1000 association matrix. Therefore, the expected number of associations per query using the full matrix would be considerably larger than the observed "average number of associations per query." Similarly, the overlap ratios would all have to be corrected by some factor in extending these results to the full matrix. However, we expect that the relative magnitudes for the various lists would not be significantly different for the vocabulary as a whole than for the subset studied.

From these results, we infer that given a query for a meaningful subject heading, the topmost portion of the sum of the association profiles of the individual words in the subject heading is almost guaranteed to contain at least a few words with meanings pertinent to the concept represented by the ordered word string (the heading). This appears to hold regardless of whether or not the heading is a System CBU or whether or not the word string occurs in our corpus. However, the expected number of such pertinent associations is least when the subject heading is an accidentally meaningful result of randomly

pairing and is most when the subject heading is selected with care* and/or is a System CBU.

5. Performance Characteristics Curves -- Subject-Heading Queries

a. Issues Investigated

The results of the tests described above and several other minor experimental forays led us to certain expectations as to the conditions under which coordinate retrieval is an adequate searching procedure, given a subject-heading query. We wished to verify these expectations directly by comparing associative and coordinate retrieval searching performance characteristics curves. Specifically, we were concerned with the question: When the query string is a System CBU, is fully automatic associative retrieval significantly different from or better than coordinate retrieval?

b. Method of Analysis

Several dozen subject-heading queries have been processed as search requests, using the "Fully Automatic Associative" retrieval search option. The procedure involved (a) inputting the subject heading to the Phase I (word association) program with equal weights on all words, and (b) using the topmost 20 associated words as input to the Phase II (message retrieval) program. The association profile for a typical query, "Surface Strain" is exhibited in Table VI-2. For this particular study, we focused specifically on a dozen requests, all of which were System CBUs. Seven of these were two-word units with $f_{ab} \geq 3$ and $C_{ab} \geq 20$, two were three-word units with $f_{abc} \geq 3$, and three were one-word units with $f_a \geq 56$.

For each of the dozen queries investigated, the retrieved messages with topmost rank were evaluated for relevance. Between 36 and 50 messages were evaluated for each query where sample size was chosen according to the density of relevant material. A scale of 5 values was chosen somewhat arbitrarily, with interpretations: 0 = not at all relevant; 1 = either vaguely relevant or minor aspects relevant; 2 = partially relevant, potentially what is wanted;

* That is, a pair of words is combined to form a concept-bearing word pair that is neither frivolous nor obscure when interpreted in the context of aerospace engineering. These are the kinds of combinations a reasonable requestor unfamiliar with the searching system might construct to represent his information needs.

JULY12	BATCH	14	ASSOCIATIVE RETRIEVAL	PAGE 1
WEIGHT	WORD	TERM NUMBER		
.9774+	STRAIN	863*		
.6767+	SURFACE	882*		
.2830+	GAGE	386		
.2422+	STRESS	866		
.1739+	PLASTIC	654		
.1352+	DEFORMATION	216		
.1246+	ELASTIC	285		
.1176+	TENSILE	898		
.1112+	CURVE	205		
.1091+	SUBJECTED	872		
.1091+	RATE	725		
.1070+	CREEP	195		
.1070+	HARDENING	411		
.0985+	LAYER	494		
.0985+	TENSION	899		
.0964+	ALUMINUM	28		
.0943+	BOUNDARY	89		
.0943+	ELEVATED	293		
.0922+	AGING	19		
.0922+	SPECIMEN	837		
(truncated)				
(Note: * means that the word was used in the request.)				

Association List Generated for the Query "Surface Strain"

TABLE VI-2

3 = most aspects relevant, almost what is wanted;
4 = very relevant, precisely what is desired. The evaluator had to be alert to all the potential meanings of the heading, for no other query context is provided in a subject heading search. We therefore found it desirable to adjust our criteria for relevance during the course of actually reading, evaluating and re-evaluating the retrieved information, with some judgments being modified as the evaluator calibrated his scale of relevance. After several passes through the retrieved items, the evaluator recorded what he meant by the various scorings for each subject heading; for example:

"Tungsten -- 4's and 3's were given to those messages concerned exclusively with the metal; 2's to those which mentioned the metal along with others, and 1's to those which mentioned "tungsten arc welding" as a minor aspect of the message's content."

For each query studied, four performance characteristic curves were plotted on a single graph; those for "Surface Strain" are shown in Table VI-3. The curve marked B is for fully automatic associative retrieval; it exhibits cumulative number of relevance points assigned to messages as a function of rank of retrieval. The curves marked A and D are, respectively, for "perfect" or "worst" retrieval for the sample studied. That is, curve A shows the result of retrieving all the 4's first, then the 3's, 2's, 1's, and finally the 0's. Curve D shows the opposite, with the 0's being retrieved first, then the 1's, etc. Curve C, "coordinate retrieval," exhibits the results of coordinate retrieval for the sample; i.e., first all messages containing both terms "surface" and "strain" are listed in their accession number sequence, and then messages containing only one term, again in accession number sequence.

The shapes of the various curves varied from request to request. Associative retrieval appears to be slightly superior to coordinate retrieval for the example exhibited in Table VI-3.

c. Results

We averaged all twelve curves at the ranks indicated to obtain the curves shown in Table VI-4. The result we expected in the first place is evident in Table VI-4; for subject-heading requests which are System CBUs there is no observable advantage to fully automatic coordinate retrieval searching. Notice, however, that the form of coordinate searching used in the comparison is not simple logical coordination of the words that make up the CBU. Rather, the entire set of messages containing any of the request terms is considered to be retrieved, arranged with those that contain both terms at the head of the list. The option to perform this type of coordinate search is becoming increasingly prevalent in computer-based systems.

Had Table VI-4 been constructed with logical coordinate search as the system against which the associative search is compared, the comparison curve would not extend beyond its position at about rank 16. The average number of messages containing both request words is about this figure. Thus a considerable amount of relevant information is gained by including the messages that contain only one of the

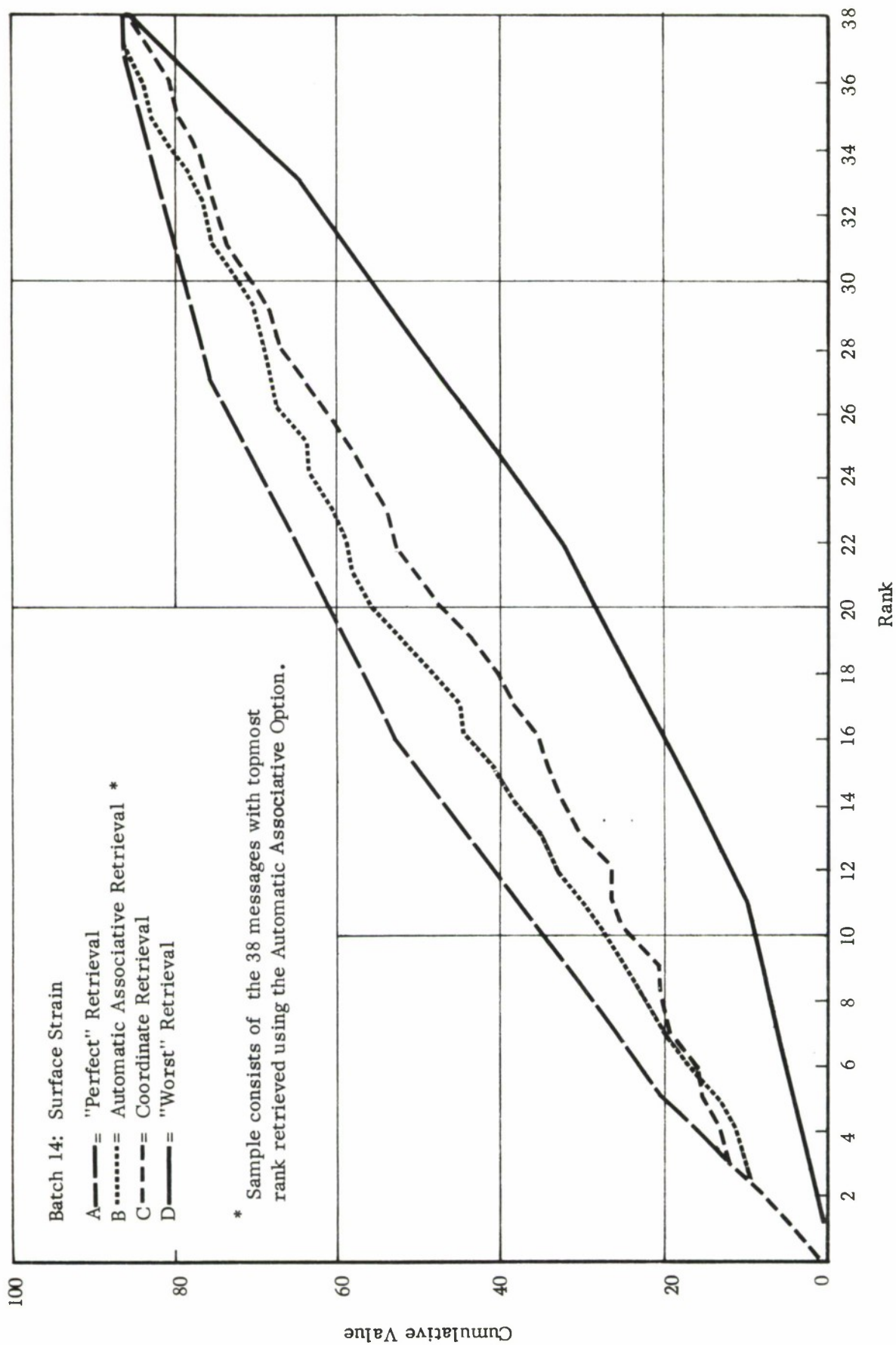


TABLE VI-3 PERFORMANCE CHARACTERISTIC CURVES FOR THE SUBJECT HEADING QUERY "SURFACE STRAIN"

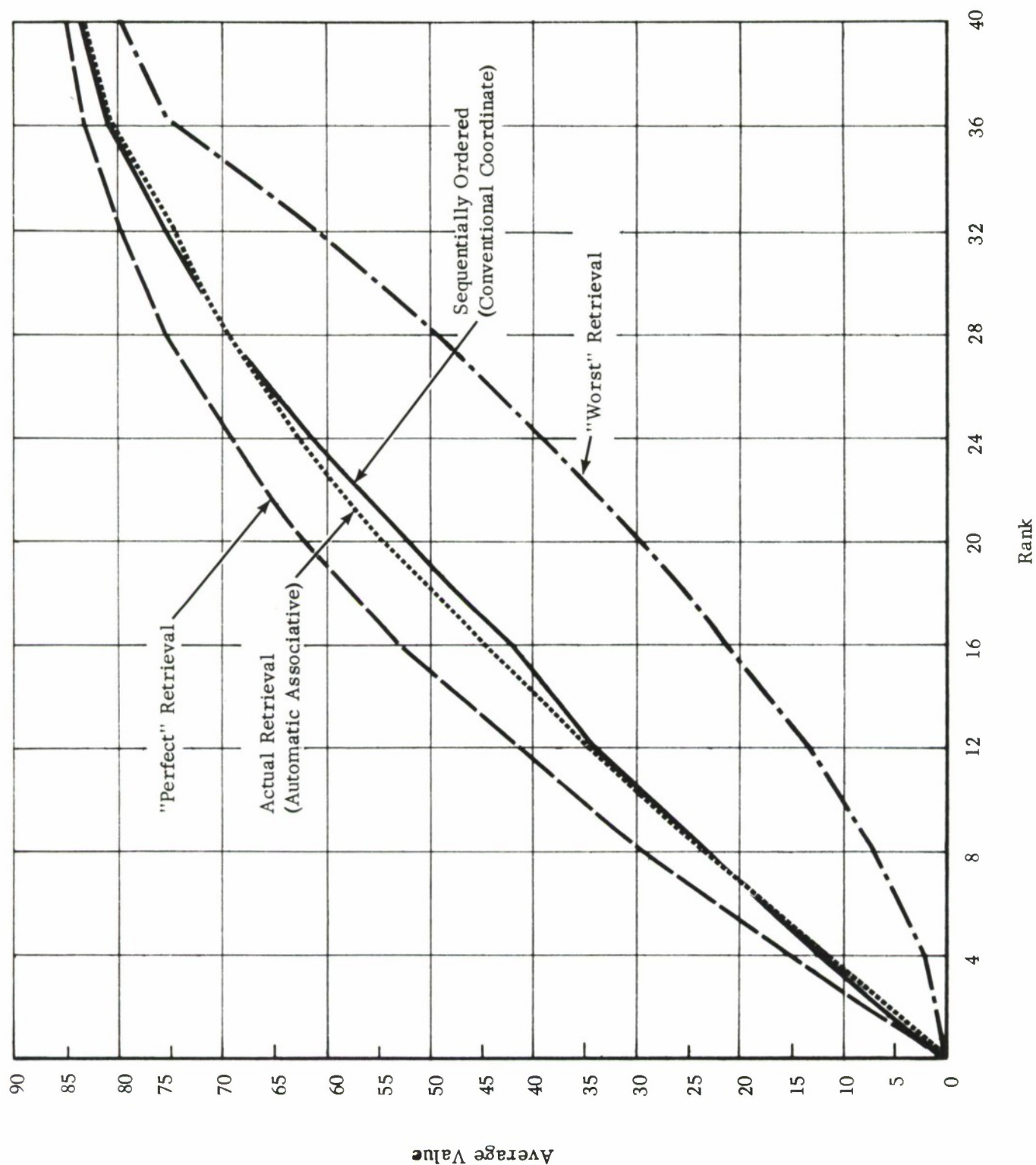


TABLE VI-4 COMPARISON OF "CONVENTIONAL COORDINATE" AND "AUTOMATIC ASSOCIATIVE" RETRIEVAL FOR QUERIES WHICH ARE SYSTEM C.B.U.'s -- Average Results for Twelve Queries

request words. Indeed, for the kinds of searches plotted in Table VI-4, there are just as many relevance points among the messages containing only one word as there are in messages containing both. (We do not yet know whether this effect is peculiar to subject-heading searches which are System CBUs. If it is generally true, one concludes that modified coordinate searching is generally a better strategy than logical search for the types of subject-heading requests considered in this section. If it is true for System CBUs only, the distinctiveness of CBU searches would be further emphasized.)

But regardless of the lessons that might be learned from further analysis, the performance curve in Table VI-4 makes it quite clear that for System CBU searches the apparatus of statistical association does not significantly improve performance over that achievable by the coordinate option. Ranking the messages by the number of query words they contain (for those searches) is a good strategy; the result is a high precision search with the recall of enough relevant material to provide a reasonable expectation that material bearing on the requestor's implicit information need will be supplied.

Without suggesting that the performance achieved on these System CBU searches is ideal, we nevertheless consider it to be well beyond the level that characterizes "acceptable" performance. Speaking subjectively, for this class of queries -- a very restricted class, by the way -- the retrieved system does respond in the fashion one would expect it ought to. It provides an interesting output that is worth looking at. And it provides little material that is distractingly irrelevant.

We emphasize this point -- that performance on these searches is acceptable -- in part to calibrate the somewhat sterile objective measures presented throughout this report. Indeed, we believe that the performance on System CBU Subject Heading searches (for the conditions of automatic indexing, short message, etc., considered in this report) is close to the best achievable performance. This is reflected by the closeness of the performance curves in Table VI-4 to the "perfect" curves. The order of presentation of the retrieval messages could perhaps be altered to produce a closer approximation to the "perfect" ordering, but we doubt that a user would notice the improvement for this class of queries. As the output stands, virtually every message retrieved in these subject-heading searches has something to do with the concept identified in the request. Because of this, we

would expect the list of messages to hold the requestor's interest as he proceeds from message to message; even though the ranking is not perfect, it is good enough not to need improving.

Accordingly, we conclude:

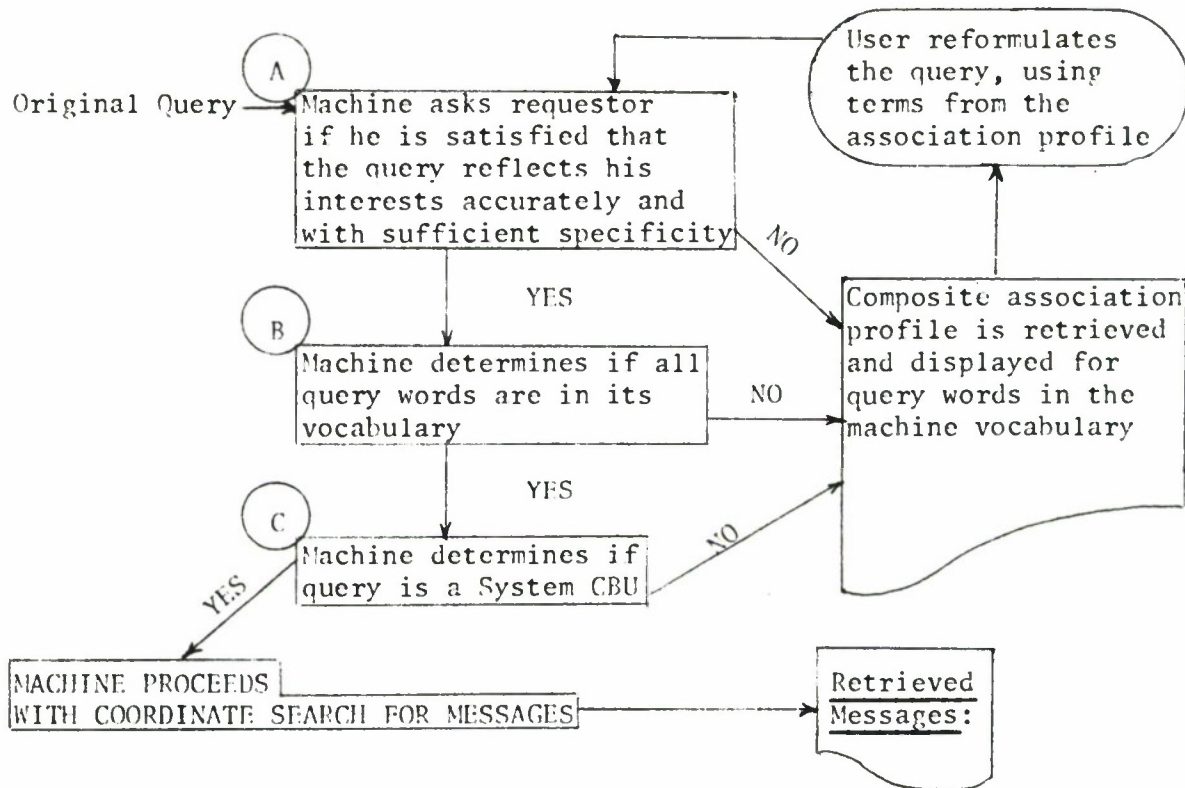
- (1) The requestor who has come upon a System CBU that represents his interest is well-advised to conduct the search and examine the messages retrieved.
- (2) We conclude also that this search need not use the associative search option but can be performed directly using the modified coordinate option (at least this is true in our collection).
- (3) When, however, the requestor has formulated a subject-heading search which is not a System CBU or one which does not reflect his interest with sufficient precision, request reformulation is advisable, for the unusual retrieval performance described in this section cannot be expected. We conclude that under these circumstances it is a better strategy for the requestor to look for one of the System CBUs that reflects his interest. It is in this process that the computed association lists play a significant role.

6. Summary-Retrieval for Subject-Heading Requests

From all the evidence available, it appears to us that the best way to process subject-heading requests with a system like our prototype is to incorporate the decision procedure of the following diagram in the computer. (Naturally, the requestor is allowed to override each of the decisions outlined.)

The resultant output of retrieved messages will have a high expected search precision (0.84 by our data) and a guaranteed recall of a minimum number of relevant messages (3, for our definition of System CBUs). The probability that the search can be done without query reformulation -- assuming the NASA STAR query statistics and assuming the requestor knows exactly what he wants -- is 0.33. To put it differently, the probability that it will be necessary to use associations and reformulate the query at least once is 0.67.

So far, we have not accumulated systematic experimental evidence as to how much effort it takes to reformulate a query using an association list, or as to how the machine can best be used to aid in this process. We do not know, for example, whether it is better to print out an association profile of single words, or actually to exhibit a



selection of System CBUs which are associated with the query. However, almost all the association lists we have observed seem to offer an immense potential for formulation of alternative subject-heading queries, whether they be lists of single words or of System CBUs. For example, from the short association list for "Surface Strain" exhibited in Table VI-2, it is possible to assemble at least 15 fairly synonymous subject headings by picking "Surface," "Layer," or "Boundary" as the first word and "Strain," "Stress," "Deformation," "Creep," or "Tension" as the second word.

It is important to observe that any list of related words is likely to be helpful as an aid to suggesting possible request reformulations. To serve this rather elementary purpose, the list does not need to be entirely "pure" (e.g., in the sense of containing only very closely related words) nor organized and formatted in intricate ways. The mere presentation of a set of words some of which are pertinent and related is already helpful. The user's job is simplified if he is presented with a choice among alternatives, whether they are offered in a book or exhibited on a console display. To serve as an aid in request reformulation, the only requirement that the list really must meet is to be rich in candidate alternatives.

If anything, the association lists provided by the prototype system are too rich in interesting suggestions for alternative query formulations; if we were to encapsulate our experience with looking at

these lists, we would have to say that alternative combinations leap from the page at any motivated requestor who is seriously thinking, at the time he inspects the list, about the given subject heading. We believe it is safe to say that the requestor almost invariably finds an acceptable alternative subject heading by combining words in the proffered list. He has little difficulty in doing so, for he rapidly perceives the interesting combinations, ignores the others, and is in a position to recycle almost at once.

The main problem that remains undiscussed in connection with the processing of subject-heading queries is determination of how quickly the requestor can be expected to "home in" on a CBU search formulation in the interactive situation being described, a high precision search with substantial recall being thereby assured.

The data we have gathered and presented earlier in this section can be used to estimate the rapidity of convergence for a subject-heading search in the process flowcharted earlier. To this end we ask, "How many times (n) does the requestor need to view the association list in order to come upon a System CBU?" We assume that the NASA subject headings we have studied are representative of the word pair combinations which the requestor will generate or recognize as subject-heading queries and use as such.

At first encounter, we assume the requestor enters the system with one of the NASA pairs that seems to him to be a good starting point. The probability that he has chosen a System CBU to start with is 0.33. This is the probability of success for $n = 0$.

In the event it is not a System CBU, he would (in the absence of overrides) be presented with an association list. With probability $(1 - 0.33) = 0.67$, then, the requestor is expected to examine the association list. In doing so, as we have argued earlier, he is nearly certain to find an acceptable alternate subject heading reformulation by combining the proffered words. The probability, however, that this pair of words is a System CBU (given that it is a subject heading and has both constituents in the vocabulary) is 0.6 for the NASA pairs. Using this figure as an estimate of the likelihood that a System CBU is found, we obtain $\text{prob. } (n \leq 1) = 1 - (1 - 0.33)(1 - 0.60) = 0.73$ and, in general,

<u>n</u>	<u>Probability of finding a System CBU subject heading search in n or fewer iterations</u>
0	0.33
1	0.73
2	0.90
3	0.96
4	0.99

For example, for 90% of the queries, two or fewer reformulations are required. These estimates are, as we have stated, based on observations supplemented by several assumptions with intuitive justifications; they may well be wrong in relative or absolute magnitude. Nonetheless, even if the single-step probabilities of success were considerably lower than we have estimated, the 0.90 figure could still be achieved in three steps of interaction instead of two. It therefore seems clear to us that the associative search-reformulation strategy outlined above can be used to obtain high performance searching using a system of the type we are considering with relatively modest investment of effort on the information searcher's part.

B. Retrieving on Full-Text Queries

We noted earlier in this report that for certain retrieval applications -- such as the correlation of contents of intelligence messages -- many new messages may be posed to the system as a query before being added to the data base.

In this application, as well as when requests are received by mail, the processing of longer queries in full-text form is indicated. We have therefore been concerned with the potentialities of a system such as ours for responding to queries expressed in full-text form, explicitly those of "paragraph" size. A paragraph of text is both an interesting size and a practical unit for the sorts of full-text queries that might be posed to a system in practice.

One population of full-text queries of the above kind is represented by the GE-0 collection of 45,000 abstracts. (Simply prefacing each abstract with "What do we have on this:" or some such expression converts it into a query.) The properties of the GE-0 collection in relation to the GE-2 machine vocabulary have been discussed extensively in Section V. To recapitulate, an average abstract from GE-0 will have an expected intersection of 16.7 terms with the GE-2 machine recognizable vocabulary. Distributions of the size of this intersection will roughly be as in Table V-2.

The main question we have been concerned with is the following: For paragraph-long queries dealing with topics covered by our collection, what is the relative effect of using different search options; in particular, how do the associative options compare with the conventional ones? Because of the large amount of effort required in making relevance judgments, we have had to choose between two alternatives: superficial analysis of the results of posing many different queries and in-depth analysis of the results of posing only a few queries. We have discussed this dilemma in Section III. We felt that the second course of action would produce the most insight during the exploratory phases of our work. We therefore have focused on retrieval experiments utilizing only a small number of full-text queries. Although we have worked in one way or the other with about

a dozen full-text queries, we focus our discussion here only on systematic tests made with four such queries and on the processing and analyzing of these same four queries in various ways. In Subsection 1 below, we treat the question of whether these queries are "typical" or representative with respect to expected query usage in connection with a collection such as ours. In Subsection 2, we describe the search options tested. Finally, in Subsections 3 and 4, we deal with various aspects of comparison of performance of retrieval search options.

1. Discussion of Our Test Questions

Our four queries consist of titles and full abstracts of items appearing in the "Propulsions Systems Section" of Technical Abstracts Bulletin, March 1, 1962 issue. The initial choice was conditioned by the subject heading of this section and the date of the issue, which appeared to be compatible with our collection. In all other respects, the selection of these particular abstracts was arbitrary.

Table VI-5 exhibits a few statistics of GE-2A vocabulary coverage of these queries and compares these statistics with averages for the messages in the GE-2 collection. Three of the queries are rather longer than the average GE-2 message. However, the average fraction of the tokens in a message covered by the GE-2A vocabulary is 0.47 for GE-2 messages and 0.46 for the test queries; the corresponding fractions for types are 0.52 and 0.51. In short, insofar as these vocabulary statistics are concerned, the queries are not different from messages comprising the collection used for retrieval.

The test queries consist of abstracts -- i.e., successions of declarative sentences. But real queries are generally expressed in interrogative form, and we were interested in whether converting each sentence of these abstracts into interrogative form would produce any observable difference on the machine processing of them. We found that we could easily and naturally rewrite each query in interrogative form, introducing vocabulary changes which amount only to the addition of words not in the GE-2A vocabulary and making a few word-order changes. (Table VI-6A shows one of our queries and VI-6B the same query rewritten in interrogative form.*) The declarative and interrogative statements are indistinguishable to our processing system because the words (those in parentheses) which make the difference are not present in the machine vocabulary -- most of these are "function words" which were deliberately excluded from our machine vocabulary.

To summarize, four full-text test queries were selected arbitrarily from an abstract journal. Vocabulary coverage of these

* Queries and their interrogative transforms are exhibited in Technical Note CACL-32, (31).

	Avg. for GE-2 Coll. Tokens Types	Q1 (Q 20) Tokens Types	Q2 (Q 02) Tokens Types	Q3 (Q 04) Tokens Types	Q4 (Q 16) Tokens Types	Avg. for 4 Queries Tokens Types
No. of Tokens (Types) in GE-2A Vocabulary	20.9 16.7	21 20	38 31	73 51	40 31	43.0 33.3
Proportion of Total Tokens (Types) in GE-2A Vocabulary	0.47 0.52	0.42 0.49	0.37 0.42	0.48 0.52	0.56 0.62	0.46 0.51

GE-2A Vocabulary Coverage
(See Text)

TABLE VI-5

AFT-END CLOSURE STUDY FOR POLARIS A-3 ROCKET MOTOR CASE

"Testing of shear test specimens for determining the optimum adhesive bond between Al cluster fitting and fiberglass chamber was continued. Hydrostatic testing of the fabricated subscale chambers was conducted. Studies of the factors and theory involved in the bond stress were made. (Author)."

Original Query

TABLE VI-6A

"[Are there any] studies of the factors and theory involved in the bond stress[es] between [the] Al cluster fitting and [the] fiberglass chamber [of] the Polaris A-3 rocket motor case? [These] aft-end closure stud[ies] [should treat the] testing of shear test specimens conducted for determining the optimum adhesive bond. Hydrostatic testing of the fabricated subscale chambers [is also of some interest]."

Note: Words in brackets are the only vocabulary differences between the two versions, and they are not included in the GE-2 machine vocabulary.

Interrogative Version of Same Query

TABLE VI-6B

queries by our machine vocabulary appears to be in the proportions observed for typical messages in the data base. Insofar as machine processing is concerned, the interrogative forms of these queries are indistinguishable from the declarative forms.

2. Search Options Tested

Altogether, ten searching options were tested; each consists of a number of steps, as follows:

a. Modified Coordinate

- (1) Full text of a written query was input to the computer, but only the 999 terms in the indexing vocabulary were recognized by the machine.
- (2) Each message in the collection was automatically assigned a weight, which is equal to the number of terms shared by the query and that message.
- (3) The texts of the messages were printed out by machine in decreasing order of their weights. No machine-computed associations were used.

b. Frequency Weighted Coordinate

This is the same as (a), only the message weights were computed differently. Each message was automatically weighted by the sum of the reciprocals of the collection occurrence frequencies of the terms shared by the query and that message. This option in effect used the term frequency "normalization" employed in associative retrieval but did not in fact use machine-computed associations.

c. Fully Automatic Associative

- (1) The original query was input as in (a).
- (2) Using a linear superposition algorithm and a previously computed word association matrix, the machine retrieved a word association profile for the given query. Each term in the profile was automatically assigned a weight, which depended upon its strength of association to the query.
- (3) The n terms in that profile with topmost value were selected by machine (where n is an experimental variable), and
- (4) Each message was automatically assigned a weight equal to the sum of the weights (from the profile list).
- (5) As in other options, messages were retrieved by the computer and printed out in decreasing order of their weights.

d. Selected Associations

This option was the same as (c) except that human mediation took place at step (3). The association profile was printed out for each query, inspected, discussed by two investigators working together, and pruned by deletion of terms which did not appear to be relevant to the intent of the original query. Machine-computed weights were kept the same, as were all other steps. Approximately five to ten minutes of time were devoted to the manual step (3) for each query.

e. Reweighted Associative

This option also involved human mediation at step (3) but otherwise was like option (d). Instead of merely pruning the association profiles, terms in the profiles were re-weighted according to human judgments of their degree of relevance to the intent of the original query. Original machine-computed profile weights were discarded, and the manually-assigned weights were used as input to step (4). Again, the manual step required about five to ten minutes per query.

f., g., h., i., j. CBU Input Options

Except for the nature of the original input material, these options corresponded exactly to options (a) through (e). Instead of using the full text of a query as input to the computer, however, a human preprocessed the query by underlining what appeared to be the most crucial content-bearing expressions. No dictionaries or lists were consulted. This required between one and three minutes per query. Only the words in those expressions were used as input to the machine. (For extremely brief queries, these options are often indistinguishable from (a)-(e).)

3. Performance Characteristics

For each of the four test queries, two sets of search options were run. In the first set, options (c) and (h) (fully automatic associative, with full-text input and with CBU inputs) were run. Because the CBU input option appeared to produce superior results -- on the average -- the later set of runs was made with options (f), (g), (i), and (j) -- all with CBU inputs. Thus altogether 24 output lists (called "batches") were inspected and evaluated for relevance.

Each batch list was inspected and messages in it were evaluated for their relevance to the query. The five level scale described in A5b above was used (0 = not at all relevant; 4 = very relevant, precisely what is desired). Each batch was presented in order of decreasing

machine-computed value, and the evaluator continued judging messages until it appeared that relevant messages were only being retrieved at random. This process usually resulted in a sample of from 100 to 200 messages being evaluated per batch, with average about 140. For a given query, cross-checks were made from one batch listing to another, to insure consistent classification of messages.* To conserve machine and human evaluator time, the machine runs were terminated after about 4,000 of the 10,000 messages in the collection had been processed by the retrieval program. That is, the results described here are essentially for retrieval from the subcollection consisting of the first 4,000 messages in our GE-2 collection. Most of the retrieved messages were found to be either not relevant or only partially relevant. Very few 3's and 4's were found. A typical batch produced anywhere from 22 to 59 relevance points. An "enriched sample" for a given query, as described in Section III, was made up of the union of the evaluated topmost portions of the output lists of the six batches for that query. Such enriched samples contain about 300 messages each and typically account for from 50 to 100 relevance points.

For each question, we plotted various performance characteristic curves, comparing the six search options in different ways. Some of the curves are exhibited in Technical Note CACL-23. Also, for each of the six options we plotted average curves of the performance for the four queries. A set of summary curves are exhibited in Table VI-7 and provide reference data for our observations comparing search options.

Each curve in Table VI-7 represents a different search option and is obtained by averaging over the four specific queries using that option. Each point on a curve shows cumulative relevance score as a function of rank of retrieval. For example, the Selected Associations option yields an average total of 18 relevance points for the first 20 messages inspected. In general, the higher a curve is, the better is the performance observed. The cutoff of 140 messages shown on the curve is arbitrary, but each tends to flatten beyond this point.

The curves of Table VI-7 exhibit several quite interesting features. Most strikingly of all, the Selected Associations option (i) outperforms either of the Coordinate options (f) or (g) by a factor of about three, and each of the other associative options outperforms the Modified Coordinate options by factors which vary from between 1.5 to 3. These factors appear to hold over the entire range of from 1 to 140 documents. That is, regardless of whether one looks at the first ten retrieved documents, the first 50, or the first 100, the listings produced by the Selected Associations options contain on the average three times as much relevant material as the listings produced by either of the Coordinate Retrieval options.

* Issues relating to single vs. multiple evaluators are discussed in Subsection C of this chapter.

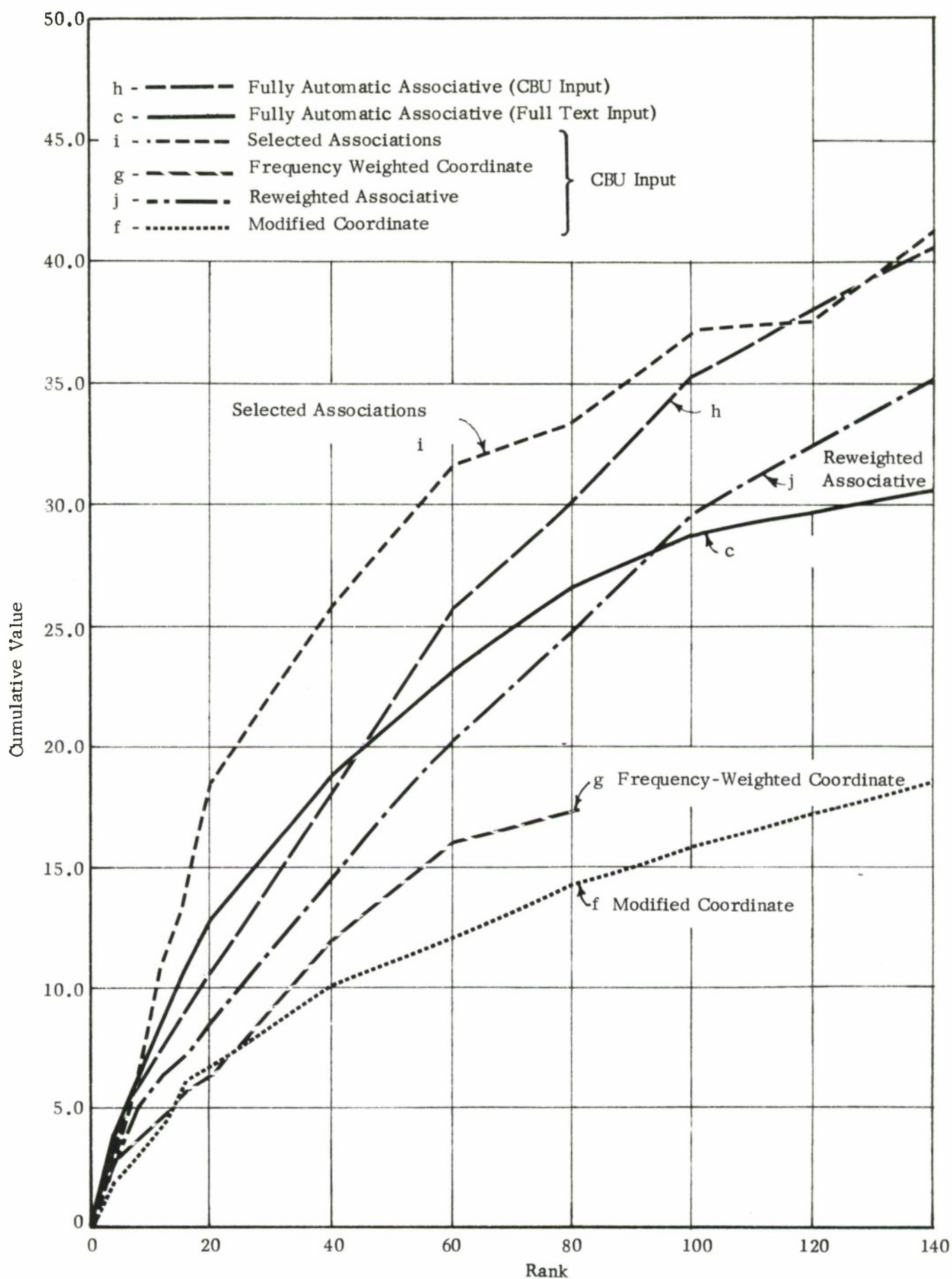


TABLE VI-7 AVERAGE PERFORMANCE CHARACTERISTIC CURVES FOR THE SIX TYPES OF REQUESTS OVER THE FOUR QUESTIONS

The next most significant observation is that the manual selection of associations appears to pay off well. This can be seen by comparing curves (i) and (h). The writers feel that the 5-10 minutes they devoted to the selection task (including card handling) could probably be reduced to 1-3 minutes, given an automated on-line display.

Likewise, looking at the two noninteractive Automatic Associative curves (h) and (c), it is clear that a considerable improvement in performance was obtained simply through CBU input; i.e., through spending a minute or two in prefiltering the input and deleting words of peripheral relevance.

Another observation has to do with human versus machine weighting of terms. Looking at the Selective Associations (i), Automatic Associative (h), and Reweighted Associative (j) curves, it seems that the machine was better at assigning weights to retrieval terms -- using frequency based criteria -- than were the humans. Reweighting association terms is not only extra bother, but appears actually to degrade performance from that the machine can do strictly on its own. If the human is to intervene effectively in the search process, it appears that the best thing he can do is simply delete unwanted association terms, leaving unchanged the machine-computed weights of those selected.

Another interesting feature seen in Table VI-7 is that the Frequency-Weighted Coordinate retrieval option (g) seems to outperform the Modified Coordinate retrieval option (f) -- not in total number of relevance points retrieved but in that it retrieves them sooner. This suggests that a modest amount of improvement in performance may be realizable in conventional systems, simply by weighting terms in inverse proportion to their occurrence frequency and using these weights in place of binary "ones" in the conventional search logic formulas.

It should be noted that the curves in Table VI-7 are averages over the four queries. The dominance of the associative over the coordinate retrieval options holds for each of the individual queries. Also, for each of the individual queries, the Selected Associations option appears to dominate the others over most but not all of the effective samples. However, each of the queries exhibits its own hierarchy of dominance relations in the various other association options. While we feel that our general observations based on these four queries will continue to be valid for full-text queries in general, we recognize that a great

deal more data will have to be processed to obtain definitive curves. A rule which appears to hold with great firmness is the following: one can predict expected performance of a retrieval system under given conditions; one can assign probabilities to various outcomes -- be prepared, however, to function in the absence of certainty, for each retrieval strategy can be foiled -- sometimes in part, sometimes in whole -- by a selected counterexample.

4. Recall Effectiveness

The tests described in Subsection B3 above left one kind of question unanswered; namely, what percentage of the relevant material in the collection was included in any of the message sample evaluated? To get at this question, we did an intensive random sampling of messages from the collection, first picking every fortieth message (by serial number) in the random sample. Later, we enlarged the sample to include every twentieth message and then again enlarged it to finally include every tenth message. Since we were dealing with only the first 4,000 messages in the collection, the final random sample had 400 messages in it. Results were essentially the same for the 400-message sample and the smaller 200-message sample. Each of these messages was evaluated with respect to each of the four queries, using the usual criteria and scaling system. Extrapolating the results, we were able to obtain estimates of the total number of relevance points in the collection (i.e., the first 4,000 messages) for each of the four queries.*

Using these estimates, we obtained estimates of "point recall" (the fraction of the total points available that was retrieved) for the four queries. These are listed in Table VI-8. The better performing search options ["Selected Associations," Option (i) or Option (h)] are seen to have retrieved samples with estimated recall between 0.4 and 0.83, with average about 0.6.

We note that the best of the six search options ("Selected Associations") shows on the average better precision than the other options over the entire range from 1 to 140 retrieved messages, and better absolute recall as well. We must reject the notion, touched on in Section IV and well publicized in the literature, that "as recall

* The 10% sample appeared to be adequate for our purposes; a discussion of error bounds in our estimates is given in a Technical Note in preparation.

	Q1	Q2	Q3	Q4
Estimated Points in 4,000-Message Collection	80	30	80	140
Points in Sample of 157 Messages for "Fully Associative Full-Text" Options {c}	37	22	35	41
Points in Sample of 157 Messages for "Fully Associative CBU Input" Options (h)	34	25	59	56
Points in Union of Two Samples Above {c} \cup {h}	41	28	62	67
Estimated Recall {c}	0.46	0.73	0.44	0.29
Ratios {h}	0.42	0.83	0.74	0.40
{c} \cup {h}	0.51	0.93	0.78	0.48
AVERAGE RECALL RATIOS FOR FOUR QUERIES:				
{c} Associative, Full-Text Input			0.48	
{h} Associative, CBU Input			0.60	
{c} \cup {h} -----			0.67	
{i} Selected Associations (Estimate)			0.60	

Estimated Fractions of Relevance Points

TABLE VI-8

risers, so precision must fall, and vice versa," as being invalid when different association-based search options are considered.*

5. Comment on Interaction

The results described in the previous sections appear to favor human mediation in query indexing and in pruning the association list. Efficient retrieval seems to require two quite different kinds of information; call them "knowledge" and "data." The needed knowledge

* The notion appears to be valid for strictly coordinate retrieval, assuming intervention of a human intermediary in the search formulation process. See, for example, the Centralization and Documentation report (20).

is about the meaning of the request and how this meaning is reflected in various language expressions. The needed data is about the statistics of language and indexing usages in the collection at hand. The latter category of information includes statistics relating to term occurrence and co-occurrence usages in the collection. The human searcher is an expert in the first kind of information, but the machine can do much better in keeping track of the second kind of information. The human searcher can usefully bring his knowledge to bear through underlining the key expressions in the query and through selection of association terms out of a machine-formulated list. However, he has only poor knowledge of the statistics of the collection and, therefore, does best if he confines his domain of selection to the set of associated terms provided by the machine and if he uses the machine-computed weights instead of his own. The machine, on the other hand, can use term co-occurrence statistics to provide large lists of terms with relatively high probability of relevance to a query but cannot tell which of them are crucial to the intent of a requestor. The partnership between man and machine clearly contributes to high-quality performance.

C. Multi-Evaluator Tests

Relatively early in the course of our evaluation work, we began to recognize how we were severely limited in what we could do with the amount of evaluator effort available. We have discussed this topic in general terms in Section III. To even approach testing the number of different search options, query types, and specific queries we wanted to test, we had to make our available evaluator manpower go as far as possible. We had to judge a large number of different message-query pairs, and this usually implied not having several people redundantly judging a given message-query pair.

We also recognized, however, that we had to conduct some tests of consistency among different evaluators, in order to establish the reliability of the observations we have been making regarding comparison of search options on the GE-2 collection. We reasoned that, if judges did not behave fairly consistently in assigning relevance scores, then it would be necessary to use several judges to obtain trustworthy results. Moreover, under these conditions it would possibly be necessary to bring to bear techniques for "unbiasing" the scores of individual judges before these scores could be meaningfully averaged. If, on the other hand, we could show that judges tend to behave fairly consistently in assigning relevance scores independently, then for certain purposes -- say, comparing search options -- the appraisals of a single judge might be trusted. That is, if under certain conditions the expected difference among relevance scores of judges was found to be small compared with the expected difference between any one judge and machine, then the use of several judges would be unnecessary.

Throughout our work we have observed that the major problem of uncertainty in assigning relevance scores occurs when the request is relatively specific with respect to the corpus; for example, when the request is a written paragraph. When the request is general with respect to the corpus (for example, when it is a subject heading which is a System CBU), there is relatively little problem of consistency. We found that there was almost universal agreement among different evaluators that any message which contains such a CBU subject heading is relevant to the concept expressed by it. The work described below is related to consistency in evaluating highly specific full-text requests.

1. What Was Done

To investigate the question of consistency of judges for search requests and the possible need for unbiasing, we analyzed the output results of four full-text search requests using two panels of three judges each. The investigations were aimed at a very specific and limited objective to answer the question:

Given retrieval performance characteristic curves for two search options, α and β , does it matter in comparing and drawing conclusions from these curves whether they are based on evaluations of only a single judge or whether they are unbiased and averaged curves for a panel of three judges?

Essentially, two distinct queries, call them A and B, were each evaluated with respect to two search options, Fully Automatic Associative with full-text input (c) and with CBU input (h). The retrieved messages in batches Ac and Ah were then given to a panel of three evaluators and those in batches Bc and Bh to another such panel, the resultant data were analyzed in various ways to be described, and certain observations were made on the results.

The evaluators worked independently, each with a separate computer-furnished rank-ordered printout of messages in the batch under consideration. The judges assigned values on the 0-4 scale used in other tests. The analysis of the resultant data was quite extensive; and, for the interested reader, discussion of it occupies some 55 pages in Technical Note CACL-27. This analysis involved the following steps:

- (a) "Unbiasing" individual evaluators' scores (by machine) using the procedure mentioned in Section III. (That is, the machine substituted, for each value V assigned by a judge, the average of the values assigned by the other judges when the given judge assigns V. For example, suppose that one of the judges assigned the value 2 to a total of 41 documents in the list. Suppose that the

average value assigned by the other judges to those 41 documents is 2.3. The unbiasing procedure would then substitute the value 2.3 for each instance when the judge in question used the value 2.)

- (b) Computing, for each of the 1,200 messages evaluated, an average of the raw scores and an average of the corresponding unbiased scores.
- (c) Computing measures of consistency -- broken down by individual evaluators, by message value range, etc. -- both for the original scores assigned and for the unbiased ones.
- (d) Preparing and comparing various performance characteristic curves, for individual scores and average scores before and after unbiasing, etc.
- (e) Conducting a study of deviations from average scores, treating the machine first as separate from individual evaluators and second as if it were an additional evaluator. This study included comparison of human and machine deviations with those expected due to a random process.

2. Results

All of the comparisons performed lead essentially to the same main result, which is independently valid for both sets of batches processed, Ac and Ah, and Bc and Bh. This is:

For purposes of comparing retrieval performance curves for two or more search options, it does not appear to matter much whether the curves are for any one of the single judges, whether they are the averaged curves for a panel of three judges, or whether, in any of the above combinations, they are unbiased. The differences are primarily ones of scale, and the relative positions of the curves for the different search options tend to be the same in all cases.

D. Manual vs. Automatic Indexing

In Section V of this report we presented data that compared the index sets assigned to messages by two indexing processes (manual and automatic), and we showed in detail how the two index sets resembled each other. We also showed that there were systematic differences

between the outputs (index sets) of the two procedures: mainly (a) the presence of spurious terms in the manual indexing, and a radical difference in the size of the indexing vocabularies employed, the manual set being some five times larger than the automatic one.

Throughout our work we have been interested in the relationship that exists between indexing "quality" and performance effectiveness. An automatic indexing of a message collection is, after all, devoid of intellect: the mere underlining of selected words in the message. Manual indexing, in contrast, involves intellectual judgment and interpretation of content to some extent; and, in choosing the set of terms to be assigned to a document, indexers apply -- consciously or subconsciously -- whatever skill and artistry their study and experience has taught them. These choices are reflected in the product of their efforts; namely, the set of index tags they decided to assign to a particular message. Since an automatic indexing has none of this, it is natural to wonder how much the intellectual part of indexing contributes to retrieval performance.

Do the two indexings yield performance of the same quality? Is the manual indexing slightly superior? Are there serious and significant differences in performance due to the difference in indexing procedure? How is the difference in performance related to the search strategy employed?

These questions could be answered at least in part by doing the following:

Given some particular retrieval methodology (e.g., one of the options of coordinate, associative, or interactive search we have discussed) and given the same collection indexed by the two procedures (manual and automatic) we have been concerned with, determine the performance differences attributable to indexing by processing the same query in what can be called a "parallel search" experiment. This experiment consists in preparing two complete retrieval systems that differ only in the message index sets used to characterize the retrievable items; everything else -- the message collection, the computational procedures, the intervention strategy, etc. -- are held constant. After the query is processed to produce a message output list, we apply the evaluative procedures to determine which system (i.e., which indexing) produced the better result and attribute the differences to differences in indexing.

We have not conducted such experiments because our collection turned out not to be suitable. That is, it was possible, by merely reasoning about the procedures to be employed, to identify residual uncertainties that would make it impossible to draw firm conclusions

from the result. Because this process of reasoning illustrates some of the considerations involved in extrapolating results from one experimental situation to another, we briefly review the problems which would be involved in conducting a parallel search experiment.

In any parallel search experiment two or more ranked output lists of messages are formed for each query, with the messages drawn from the same collection. The procedure we employed in comparing search options earlier in this chapter was exactly of this kind. To compare the results, one would form the union of all messages retrieved by the competing systems, evaluate the messages for relevance by assigning points, and plot the comparative performance curves. Any systematic tendency of one system to place more relevant messages near the top of the list would be apparent in the relative position of the competing performance characteristic curves.

The discussion in Section V indicated that the topical content of messages was covered to a comparable extent by the two indexing processes (i.e., machine and manual UNITERM indexing) performed on them. However, for a typical message, several of the assigned UNITERMS correspond to topics which are simply not mentioned in the text of the message itself. We conjectured that this was due to the fact that the indexer was trying to represent the parent document rather than the abstract and noted that it was possible to identify two observably different subpopulations -- one presumably of author-generated abstracts, the other presumably of indexer-generated abstracts.

Whatever the explanation for the spurious terms, we recognized that their presence posed a serious obstacle to retrieval evaluation using the manual indexing. For, in those cases where the manually-assigned index set is designed to represent the parent document, we could not in fairness judge relevance simply by reading the text of the abstract. Consider an abstract that is retrieved by virtue of hits on one or more spurious terms; when the evaluator reads the abstract, this document is likely to seem less relevant than the parent document really is (if the indexer was right). The evaluator will tend to underestimate its relevance, and thus an experimental design that bases the development of a master list on abstracts would systematically prejudice the evaluation against the manual indexing. On the other hand, since the automatic indexing has no way of detecting the content of the parent document, basing the evaluation on parent documents could systematically prejudice the master list against the automatic indexing. The evidence (inferred from the presence of spurious terms) that the manually-assigned terms do not necessarily apply to the abstract thus made it impossible to use our data directly for a parallel search experiment. There was no clear-cut way to evaluate the output impartially, and the objective of a pure experiment to compare the effects of indexing had to be abandoned.

We also considered other, less rigorous, avenues for achieving substantially the same sort of comparison. One alternative was to attempt to isolate the "Population R" abstracts discussed in Section V,

i.e., to find documents for which the UNITERM set does index the abstract (or at least can be assumed to do so by virtue of having few spurious assignments). Unfortunately, no systematic criterion for accomplishing the selection was available in our collection. Another alternative was to delete the spurious term assignments in a substantial subset of the documents in the collection with the idea that the remaining UNITERMS would characterize the abstract, but the result of this deletion step would have been UNITERM sets that were not the product of an actual indexer's decision, a factor which would have obscured further the conclusions one might have drawn from the experiment.

Accordingly, we found that reasonable alternative procedures were not available for carrying out an experimental comparison of the performance differences attributable to indexing in our collection.

Nevertheless, the indexing comparison studies described in Section V lead us to expect or conjecture the existence of the following relative performance differences.

Suppose the parallel search experiment had been conducted with the given data.

a. Logical Coordinate Searching -- Manual Indexing
vs. Machine Indexing

Consider a search request which consists of a single term or of a logical product of several terms, all of which are common to the GE-0 and GE-2 vocabularies; moreover, assume that retrieval is of the 45,000 abstracts in the GE-0 parent collection. Then, on the average, we expect that this search will retrieve roughly the same number of relevant messages using either indexing -- i.e., the recalls of the two searches will be comparable (because term usage frequencies are comparably correlated and because messages are indexed to roughly the same conceptual depth). However, the expected precision of the search using the manual GE-0 indexing will be worse than that of the machine GE-2 indexing because of the spurious manual term assignments.

b. Weighted Coordinate Searching -- Manual
vs. Machine Indexing

Suppose that search request and collection are as in (a), but the search algorithm weights all retrieved messages according to the number of request terms they contain. Then, the expected result is that the performance characteristic curve (exhibiting the total

number of relevant retrieved documents as a function of their rank in retrieval) will be lower for manual GE-0 indexing than for GE-2 indexing, again because of the spurious manual term assignments.

c. Associative Searching -- Manual vs. Machine Indexing

If the human effort connected with searching (devoted to selecting associations, etc.) is kept constant, there is no reason to believe that the comparison stated in (b) will not continue to hold in the associative case. That is, a more "noisy" indexing will lead to lower performance whether or not associations are used.

d. Manual Indexing -- Coordinate vs. Associative Searching

We observed that the manual GE-0 indexing appears to give coverage of concepts in abstracts comparable to GE-2, but that the UNITERM GE-0 indexing includes a number of spurious terms. Other than for these spurious terms, the UNITERM and machine indexing parameters of a typical message are nearly identical. It therefore seems reasonable to expect that the spurious indexing terms would tend to penalize performance of the manually-indexed system regardless of whether a coordinate or associative search option is employed. Thus, we would expect that the relative differences between performance of various search options observed in GE-2 and discussed previously in this chapter would also be present using the original GE-0 manual indexing, with over-all performance being degraded somewhat. This difference, it will be recalled, is that associative searching with manual mediation does about three times as well as simple coordinate searching for full-text queries.

It has not been our objective, in the course of this laboratory evaluation, to consider comparisons of processing economics of various procedures. It is very clear to us, however, that the automatic indexing choice would be apt to enjoy major processing cost advantages, given any new operational situation; namely,

e. Probable Processing Economics

Our guess, based mainly on our own experience and on studies with which we are familiar, is that cost of automatic indexing of a large collection (100,000 or more messages) including costs of message transcription (but

not, obviously, initial message preparation) should be cheaper than manual indexing by a factor of between five and ten. Also, there are additional processing economics to be achieved through using a much smaller index term set (999 terms vs. 4,700 for comparable coverage in our case) to accomplish comparable conceptual depth of coverage of messages. In particular, if the system is to be completely associative, i.e., use machine-computed associations among all index terms, then an automatically-indexed collection can be expected to require processing of a much smaller matrix (by a factor of about 22 for our collection) than a manually-indexed one.

Two words of caution are essential. First, our points (a)-(e) above are not established by experiment; they represent our best appraisal of what would be likely to happen, but one or more could be inapplicable in a given specific instance. Second, items (a)-(e) and, in fact, all of the discussions in this report which tend to show the original GE-0 manual indexing to be at a disadvantage, clearly apply only to the message retrieval application we are considering, not to the document retrieval application for which the UNITERM indexing was originally intended. We continue to assume that the existing manual indexing is serviceable and useful within the context of GE's practical application. We have not looked specifically into the question of whether or how automatic indexing would be appropriate within an environment where the main objective is the retrieval of full-length documents, although some of our results obviously bear on this issue.

REFERENCES

1. Arthur D. Little, Inc., "Automatic Message Retrieval," Report FSD-TDR-63-673 (November 1963).
2. Arthur D. Little, Inc., "Towards the Use of Natural Language Structure in Automatic Message Retrieval," an Interim Report, Technical Note CACL-18 (June 1965).
3. Taube, Mortimer, "A Note on the Pseudo-Mathematics of Relevance," American Documentation, 14, No. 2 (April 1965), p. 69.
4. Hillman, Donald J., "The Notion of Relevance," American Documentation, 15, No. 1 (January 1964), pp. 26-34.
5. Oettinger, A. G., (Ed.), "A Forum on Centralization and Documentation," Communications of the ACM, 8, No. 11 (November 1965), pp. 704-10.
6. Henderson, Madeline M., "Bibliography on Evaluation of Information Systems," National Bureau of Standards (in press).
7. Cleverdon, Cyril W., The In-House Testing and Evaluation of the Operating Efficiency of the Intellectual Stages of Information Retrieval Systems, College of Aeronautics, Cranfield, England (1964). (An amended version of a paper presented to the FIC Conference on Classification Research at Elsinore, Denmark (September 1964).)
8. Cleverdon, Cyril W. and Mills, J., "The Testing of Index Language Devices," ASLIB Proceedings, 15, No. 4 (April 1963), pp. 106-30.
9. Cleverdon, Cyril W., and Aitchison, J., A Report on a Test of the Index of Metallurgical Literature of Western Reserve University, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England (October 1963), 270 pp.
10. Cleverdon, Cyril W., "A Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England (October 1962), 305 pp.
11. Richmond, Phyllis A., "Review of the Cranfield Project," American Documentation, 14, No. 4 (October 1963), pp. 307-11.
12. Swanson, Donald R., Design of Experiment for the Testing of Indexing Systems: A Review of the Cranfield Reports (publication status unknown), University of Chicago, Library School (August 1964), 29 pp.

13. Salton, Gerard, "The Evaluation of Automatic Retrieval Procedures -- Selected Test Results Using the SMART System," American Documentation, 16, No. 3 (July 1965).
14. Dale, A. G. and Dale, N., "Some Clumping Experiments for Associative Document Retrieval," American Documentation, 16, No. 1 (January 1965), p. 5.
15. Curtice, R. and Rosenberg, V., "Optimizing Petrieval Results With Man-Machine Interaction," Center for the Information Sciences, Lehigh University (February 1965).
16. Mooers, Calvin N., The Intensive Sample Test For the Objective Evaluation of the Performance of Information Retrieval Systems, ZATOR Bulletin ZTB-132, The Zator Company, Cambridge, Mass. (1959).
17. Fels, E. M., "Evaluation of the Performance of an Information Retrieval System by Modified Mooers Plan," American Documentation, 14, No. 1 (January 1963), pp. 28-34.
18. Bornstein, Harry, "A Paradigm for a Retrieval Effectiveness Experiment," American Documentation, 12, No. 4 (October 1961), pp. 254-9.
19. Bryant, Edward C., "Progress Towards Evaluation of Information Systems," paper presented at the ICIREPAT Meeting, Washington, D.C. (October 1964), U. S. Patent Office, Office of Research and Development.
20. Arthur D. Little, Inc., Centralization and Documentation, Report to the National Science Foundation, Second Edition (June 1964); see also the Appendix to the Second Edition.
21. Altmann, Berthold, "A Multiple Testing of the ABC Method and the Development of a Second-Generation Model, Part II: Test Results and an Analysis of 'Recall Ratio'." TR-1296, U. S. Army Materiel Command, Harry Diamond Laboratories, Washington, D.C. (October 1965).
22. Cleverdon, Cyril W., Lancaster, F. W., and Mills, J., "Uncovering Some Facts of Life in Information Retrieval," Special Libraries, Vol. 55, No. 2 (February 1964), pp. 86-91.
23. Kendall, M. G., Rank Correlation Methods, Third Edition, Hafner Publishing Company, New York (1962), Chapter 2.
24. Mazuy, Kay K., "A Modification of Kendall's Rank Correlation Measure," Presented at the Annual Meeting of the Institute of Mathematical Statistics, Amherst, Mass. (August 1964).

25. Giuliano, Vincent F. and Jones, Paul E., "Properties of Corpus GE-1," Technical Note CACL-12 (March 1965).
26. Mehring, Joyce S., "Word Units, Frequency Counts, and Machine Indexing of GE-2 Data Base," Technical Note CACL-13 (March 1965).
27. Bono, Peter R. and Jones, Paul E., "NASA Vocabulary Two-Word Strings, Their Usage, and Relation to System CBUs in the GE-2 Auto-Indexed Message Collection," Technical Note CACL-31 (June 1966).
28. National Aeronautics and Space Administration, Guide to Subject Indexes for STAR, NASA SP-7016, Revision 1 (February 1965).
29. Giuliano, Vincent F. and Jones, Paul E., "Analysis of Preliminary Retrieval Experiments Using CBU Components as Search Terms," Technical Note CACL-17 (May 1965).
30. Mehring, Joyce S., "QUES: A Computer Program for On-Line Queries of Association Profiles," Technical Note CACL-26, Supplement 1 (December 1965).
31. Bono, Peter R., "Vocabulary Overlap Comparison Between Average Full-Text Query and Typical GE-2 Messages," Technical Note CACL-32 (June 1966).

APPENDIX A*

ESTABLISHING A MASTER LIST: COMBINING THE JUDGMENTS OF SEVERAL EVALUATORS

A. General Problem

In Section 111 we discussed the general nature of a "master list." For a given query and a given sample of messages, such a list assigns a presumably "true" or "master" value of relevance to each message with respect to the given query. The master list values are used as standards when comparing performance of various retrieval options.

A master list could be prepared by a single judge who, presented with a query, a sample of messages and a scale of values, is told to assign a relevance value to each message. However, it may be felt that a single individual is not adequate to represent the user population of a retrieval system, or that his judgments might well be contested by other judges. We have therefore been concerned with how several such lists of judgments made by different individuals might be reconciled and combined together to make a master list representing the over-all consensus of the judges.

In this appendix we consider two approaches to effecting such a reconciliation, an "Error Matrix" approach and a "Simple Unbiasing" approach. Both approaches embody provisions for "unbiasing" -- i.e., for taking into account the tendencies of some evaluators to systematically score higher or lower than others.

In order to discuss the mathematical aspects of such a reconciliation, the following terminology is introduced:

D = Number of messages in the sample

J = Number of judges

V = Maximum value that can be assigned by any judge

v_{ij} = Value assigned to message i by judge j

$\bar{v}_i = \{v_{i1}, v_{i2}, \dots, v_{iJ}\}$ = Value vector for message i

The general statement of the problem is practically identical to that solved by linear discriminant techniques: Given D value vectors \bar{v}_i ($i = 1, 2, \dots, D$), what adjusted value t_i should

* The material in this Appendix was contributed by S. Pollock.

be assigned to message i , such that an ordering of the messages by t_i reflects an ordering by "true value," or "relevance."

B. The Error Matrix Approach

One method by which this problem may be attacked is to assume that each judge j has some constant probability of assigning a value v_{ij} to message i , given that the message has, in fact, some "true" or real value ℓ . In particular, let the possible values v_{ij} consist only among the positive integers up to V (i.e., $v_{ij} = 1, 2, \dots, V$; $i = 1, 2, \dots, D$; $j = 1, 2, \dots, J$). Then we define the error probabilities p_{jkl} .

$$p_{jkl} = \text{probability \{judge } j \text{ assigns a value } v_{ij} = k \text{ to message } i, \text{ given that message } i \text{ has in fact the real value } \ell.\}$$

$$(i = 1, 2, \dots, D; k, \ell = 1, 2, \dots, V).$$

The j th judge's error matrix is $P_j = p_{jkl}$.

In addition, we consider the probabilities q_m ($m = 1, 2, \dots, V$), where $q_m =$ a priori probability {a message has value m }.

We have thus introduced $V(V-1)$ variables for each judge (not V^2 , since the P_j are stochastic matrices, and hence rows must sum to unity) and V variables for the a priori probabilities, a total of

$$JV(V-1) + V$$

independent variables.

If these variables were known, then by direct application of inferential probability statements, the distribution of the "true" message values, given some answer vector \bar{v}_i could be obtained. In order to determine these variables, it is possible to treat them as unknown, and use observed data to obtain best fits by standard statistical techniques.

Some study of the model presented above reveals that the sufficient statistics consist of the observed number of each of the V^J possible response vectors. All of these must add to the total number of response vectors, which is equal to the total number of messages. For significant results, then the number of degrees of freedom

$$D.F. = (V^J - 1) - (JV(V-1) + V)$$

must be reasonably large. For $J = 3$, $V = 3$,
 $D.F. = (27-1) - (3 \cdot 3 \cdot 2 + 3) = 5$, the approach is of marginal
 use.

The main difficulty with this technique as developed so
 far, however, is that the final result is in the form:

message i has probability a_{i1} of having value 1
 a_{i2} of having value 2
 \vdots
 a_{iV} of having value V

Because each message is thus characterized by a V -dimensional
 vector, a consistent ordering (by valuation) scheme is not immediately
 available. This is because the values $1, 2, 3, \dots, V$ do not necessarily
 reflect a linear relationship between the actual relevance of the
 message: a message with 0.5 chance of being value 0 and 0.5 change
 of being value 2 is not necessarily as relevant as the message that
 has probability 1 of having value 1.

We have thought of one way to get around this difficulty, and
 that is by defining continuous functions, describable on $(0,1)$ by two
 parameters, of the form

$f_r(v; \alpha_i, \beta_i) =$ probability (judge i assigns value v to
 a message, given that its true value is r)

In addition, the a priori probability of any given message
 having value v can also be regarded as a two parameter function
 $g(v; \delta, \gamma)$. The total number of variables thus introduced is $2J$ for
 the α_i and β_i , and 2 for γ and δ . Thus, the degrees of freedom are

$$D.F. = (V^J - 1) - (2J + 2)$$

and for $J = 3$, $V = 3$, $D.F. = (27-1) - (6+2) = 18$, which is certainly
 stronger than before. Since the number of degrees of freedom is
 considerably improved, higher statistical significance can be expected.

The assigning of values to each message is then straightforward:
 the result of the calculations is to provide a posterior distribution
 over v , given that a particular answer vector \underline{v}_i is observed. The
 moments of this posterior distribution (particularly the mean) will
 then indicate relative relevances of the messages although again a
 simple ordering is not assured.

In summary, an error matrix approach can provide a statistically viable avenue to creating an unbiased master list, but the method requires estimating a great many quantities and is cumbersome to use.

C. Simple Unbiasing

In order to obtain an easy-to-use technique for creating a master list, a simpler method is proposed here. This method involves two steps:

- (1) On the basis of the evaluations of all judges on all messages, develop a set of corrections for each judge that will compensate for his consistent biases.
- (2) Using each judge's corrected values, apply standard statistical techniques to obtain average values, and associated measures of inconsistencies (i.e., analysis of variance).

The corrections are in the form of a matrix $S = \{s_{jk}\}$, where s_{jk} = corrected value to be used when judge j assigns the value k ($k = 1, 2, \dots, V$).

The generation of this matrix is straightforward: take every message that judge j has assigned a value k to, determine the average value of scorings (over all judges) of the messages in this set, and use this value instead of k (for judge j). The average value may be calculated either including or excluding judge j 's evaluation. The latter is recommended when the number of judges is fairly large (above 5). Once S has been calculated, it is used to "calibrate" each judge's response. As more data are obtained, S becomes a better indicator of the judges' relative value scales.

A convenient representation of S for analytical purposes is:

$$s_{jk} = \frac{\sum_{i=1}^D \sum_{m=1}^J v_{im} \delta(v_{ij} - k)}{\sum_{i=1}^D \sum_{m=1}^J \delta(v_{ij} - k)}$$

$$\text{where } \delta(n) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \quad \text{for}$$

The summation over the judges ($m=1$ to J) may include or exclude judge i , as mentioned above.

The value t_i of document i assigned by this method thus becomes

$$t_i = \frac{1}{J} \sum_{j=1}^J \sum_{k=1}^V s_{jk} \delta(v_{ij}-k)$$

The unbiasing scheme was designed to account for the unavoidable subjective interpretations by the judges of an arbitrary 1 to v value scale. We have tested it on a limited scale and have found it to be workable (see Section VI), although we have questioned the practical necessity of any technique for unbiasing.

APPENDIX B*

MEASURES OF USER SATISFACTION WHICH RELATE TO SEARCH OBJECTIVES

We observed in Section IV that a statistic which merely compares two message lists -- say, a master list and a retrieval output list -- and decides whether or not they are "similar" in some sense, is not sufficiently revealing of system performance attributes. The degree of similarity so measured might be in just the portion of the list that is unrelated to the real use to which the list of messages will be put. In this appendix, we develop and discuss four possible measures of user satisfaction; each of these can be made sensitive to the desired "depth" of search, as discussed in Section IV. The measures dealt with are called:

- (1) Normalized Sliding Ratio Measure
- (2) A Browser's Statistic
- (3) A Weighted Rank-Correlation Statistic
- (4) Cost-Matrix Measures

A. The Normalized "Sliding Ratio" Measure

This measure is appropriate for comparing pairs of systems and, in order to provide an absolute measure, requires the existence of a master list. The measure has a relation to the classical recall and precision ratios and may be looked upon as a generalization of them. The Sliding Ratio measure fits in more naturally with the use of performance characteristics curves than do the other measures discussed in this appendix, and we therefore discuss it more thoroughly, giving examples.

1. Assumptions and Ground Rules

- (a) The collection consists of a total of N messages, any one of which is retrievable by the candidate systems.
- (b) The output of a system is in general an ordered list L . $L = \{l_1, l_2, l_3, \dots, l_N\}$ where l_i is the identifying number (serial number, etc.) of the i th message in the list.

* The material in this Appendix was contributed by S. Pollock.

- (c) This list is further broken down into K groups called ranks, ($K < N$) such that the first k_1 messages are in the first rank, the next k_2 messages are in the second rank, and so on, until the last k_K messages are in the K th rank. Since all messages must be in some rank, r ,

$$\sum_{r=1}^K k_r = N .$$

- (d) A fully-ranked list is defined to have only one message in each rank, so that all the k_j are equal to unity and the number of ranks K is equal to N .

The opposite case is the pathological one where $K = 1$, so that $k_K = k_1 = N$, and all messages are contained in the first rank.

Many retrieval systems will have $K = 2$ ranks, with k_1 selected messages included in the first rank and the remaining $N - k_1 (= k_2)$ messages in the second rank.

- (e) In order to obtain an absolute measure, a master list L^* is assumed to exist, with the properties of general lists described above, and in addition

(a) has K^* ranks with k_i^* messages in rank i ;
($i=1,2,\dots,K^*$);

(b) associates a value v_r with a message in the r th rank of the master list.

2. Definition of the "Sliding Ratio" Measure

The following is the formal definition of the measure; the discussion in the following section will justify its form and discuss its properties.

For any list L and a master list L^* , let us define

$x_{ij} \equiv$ the number of messages in rank i in L that are in rank j in L^*

and

$$\begin{cases} g_0 = 0 \\ g_t \equiv \sum_{r=1}^t k_r \quad (t=1, 2, \dots, K-1) \end{cases}$$

then a function $f(n)$ may be defined for

$$g_t < n \leq g_{t+1} \quad (t=0, 1, \dots, K-1)$$

$$f(n) = \frac{n-g_t}{k_{t+1}} \sum_{j=1}^{K^*} v_j x_{t+1,j} + \sum_{i=1}^t \sum_{j=1}^{K^*} v_j x_{ij} \quad (1)$$

When $L = L^*$ in the above definitions, then we may similarly define

$$x_{ij}^* = k_i^* \delta_{ij} \quad \delta_{ij} = \text{Kroneker delta}$$

$$g_t^* = \sum_{r=1}^t k_r^* \quad g_0^* = (t=1, 2, \dots, K^*-1)$$

and so for $g_t^* < n \leq g_{t+1}^*$

$$f^*(n) = \frac{n-g_t^*}{k_{t+1}^*} \sum_{j=1}^{K^*} v_j x_{t+1,j}^* + \sum_{i=1}^t \sum_{j=1}^{K^*} v_j x_{ij}^* \quad (2)$$

$$= (n-g_t^*) v_{t+1} + \sum_{i=1}^t v_i k_i^*$$

The "Sliding Ratio" Measure is now defined to be

$$\mu(n) = \frac{f(n)}{f^*(n)} \quad (3)$$

3. Clarification and Discussion of the Sliding Ratio Measure

This section provides a rationale for the use of the $\mu(n)$ measure just derived. Let us allow the assumption that there is a master value assigned to each message in the collection. In addition, we assume that there are K^* levels, or ranks, of this value, so that any message has the master value v_r , $r = 1, 2, \dots, K^*$. By ordering all the messages in the collection according to these values, the Master List is constructed.

One way of interpreting these values is to consider them as measures of the amount of query-related content contained in each message. When interpreted in this way, it is evident that the master value of one message is independent of the master values of the other messages in the collection. In fact, this is a highly desirable property to assign to the value scale, since the existence of other messages should not affect the "retrievability" of the one in question. With this interpretation of the master value of each message, we see that if we are presented with a specific set of messages, the master value of this set is the sum of the master values of the messages contained.

Now let us constrain the requestor to use only the first n messages of a L-list produced by a system. When the list is fully ranked, there is no question as to what is meant by the "first n" messages: simply those n messages in the first n ranks (each rank has one message). On the other hand, if the list has many messages tied at various ranks, then the concept of the "first n" messages needs clarification.

When we say the requestor may use only the first n of the list, we shall imply that if, at any point, the first n messages have not been obtained, and two or more messages have the same highest rank remaining, then one message is selected from among these, at random. For example, (3, 5, 1, 2, 4) and $k_1 = 1$, $k_2 = 1$, $k_3 = 3$, the list may be written:

Message #	3	5	1	2	4
Rank	1	2	3	3	3

Then if $n = 4$, each of the following sets of messages has equal probability of being the "first 4" of the list:

(3, 5, 1, 2)	(3, 5, 2, 4)
(3, 5, 1, 4)	(3, 5, 4, 1)
(3, 5, 2, 1)	(3, 5, 4, 2)

If $n = 2$, then the first n messages can only be (3,5).

In order to justify the $\mu(n)$ measure, we simply imagine that, in fact, the requestor must limit himself to selecting only n messages from the L list. If the L list has any ordering, this implies that these first n should be selected according to this ordering, with tied ranks being treated as discussed above.

The absolute values of the selected n messages, however, depend only upon their ordering according to L^* , the master list. Since we have assumed that the total value of the n messages selected is equal to the sum of their individual values, we are led to the form of equation (1). The double summation in the right-hand side of this equation is seen to be the sum of the values of the messages in the first t ranks of the L list. The first (single) summation represents the average value (weighted over all appropriate combinations) of messages from the $(t+1)$ st rank, needed to bring the total of messages selected to n .

If we wished, we could simply use the $f(n)$ of equation (1) as our measure. This is what we do in practice when we draw an unnormalized performance characteristics curve. However, the value of this function depends very strongly (naturally) upon the master list, and the values associated by it to each rank. In order to normalize this function, we can form the function $f^*(n)$ of equation (2), which is simply the total first- n -message value achieved by using the L^* list as the L list; in other words, $f^*(n)$ is the best you can do when required to select n messages.

The measure $\mu(n)$ then is a normalized one and will equal unity when $L = L^*$ and will always be less than unity when the L list differs from the L^* list.

4. Numerical Examples

(1) Fully-ordered Master List and System List ($N=5$)

(a) Master List L^* :

3	5	1	2	4	
rank:	1	2	3	4	5
value:	10	8	5	2	0

($K^*=5; k_1^*=k_2^*=k_3^*=k_4^*=k_5^*=1$)

(b) System List L:

3	4	5	1	2
1	2	3	4	5

 ($K=5; k_1=k_2=k_3=k_4=k_5=1$)
rank:

(c) g-calculation: $g_0 = 0; g_1 = 1; \text{etc...}$ $g_i = i \quad i = 1, 2, \dots, 5$
 $g_0^* = 0; g_1^* = 1; \text{etc...}$ $g_i^* = i \quad i = 1, 2, \dots, 5$

(d) x_{ij} Matrix

	Rank in L*				
	1	2	3	4	5
<u>Rank in L</u>	1	1	0	0	0
	2	0	0	0	1
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1

(e) $f^*(n)$ calculation
Since $g_t^* = t$, and $k_i^* = 1$, ($i=1, 2, \dots, 5$), equation (2) reduces to

$$f^*(n) = \sum_{i=1}^n v_i$$

n	1	2	3	4	5
$f^*(n)$	10	18	23	25	25

(f) $f(n)$ calculation
Since $g_t = t$, and $k_i = 1$, equation (1) reduces to

$$f(n) = \sum_{i=1}^n \sum_{j=1}^5 v_j x_{ij}$$

n	1	2	3	4	5
f(n)	10	10	18	23	25

(g) $\mu(n)$ calculation: $\mu(n) = \frac{f(n)}{f^*(n)}$

n	1	2	3	4	5
$\mu(n)$	1	0.55	0.78	0.92	1

(h) Comments: Note that if only one message is allowed (or desired) then the retrieval system is scored unity (ideal) -- the best possible system (the L^* list) would present the same message. Similarly, if the whole collection were demanded, the system again scores unity (as would even the L^* system) because no discrimination was desired. The behavior of $\mu(n)$ as n goes from 1 to N thus indicates the ability of the system, and the value of $\mu(n)$ at a particular point is not of interest by itself, unless we are ready to concede the constraint of requiring n messages to be selected.

2. Master List and System List Involving Ties ($N=5$)

(a) Master List L^* :

3	5	1	2	4	
rank:	1	1	2	2	3
values:	9	9	3	3	0

($K^*=3; k_1^*=k_2^*=2, k_3^*=1$)

(b) System List L :

3	4	5	1	2	
rank:	1	1	1	2	2

($K=2; k_1=3, k_2=2$)

(c) g -calculations: $g_0 = 0; g_1 = 3; g_2 = 5$

$$g_0^* = 0; g_1^* = 2; g_2^* = 4; g_3^* = 5$$

(d) x_{ij} Matrix Rank in L^*

		1	2	3
<u>Rank in L</u>	1	2	0	1
	2	0	2	0

(e) $f^*(n)$ calculation

$$0 < n \leq 2 \quad f^*(n) = nv_1$$

$$2 < n \leq 4 \quad f^*(n) = (n-2)v_2 + 2v_1$$

$$f^*(5) = v_5 + 2v_1 + 2v_2$$

n	1	2	3	4	5
$f^*(n)$	9	18	21	24	24

(f) $f(n)$ calculation

$$0 < n \leq 3 \quad f(n) = \frac{n}{3} \sum_{j=1}^3 v_j x_{1j}$$

$$= \frac{n}{3} [v_1 x_{11} + v_2 x_{12} + v_3 x_{13}]$$

$$= \frac{n}{3} [9(2) + 3(0) + 0(1)] = 6n$$

$$3 < n \leq 5 \quad f(n) = \frac{n-3}{2} \sum_{j=1}^3 v_j x_{2j} + 18$$

$$= \frac{n-3}{2} [9(0) + 3(2) + 0(0)] + 18 = 3n + 9$$

n	1	2	3	4	5
f(n)	6	12	18	21	24

(g) $\mu(n)$ calculation: $\mu(n) = \frac{f(n)}{f^*(n)}$

n	1	2	3	4	5
$\mu(n)$	0.67	0.67	0.86	0.88	1

- (h) Comments: If only one message were called for, then the L list has a 1/3 chance of selecting message #4, which is not first ranked in the L* list, so that the measure of this system, at $n = 1$, is 2/3.

5. Normalized Precision and Recall Ratios: $K^* = 2$

When messages can only have two degrees of relevance to a query (either "total," or "none") the master list consists of only two ranks ($K^*=2$), although it is possible for a system to be "unaware" of this restrictive dichotomy. Then it is possible to readjust the value scale (since values are additive) so that $v_1 = 1$, $v_2 = 0$: Relevant messages (rank 1) are worth one unit, irrelevant messages (rank 2) are worth nothing.

By using these restrictions, equation (1) becomes for $g_t < n \leq g_{t+1}$:

$$f(n) = \frac{n-g_t}{k_{t+1}} x_{t+1,1} + \sum_{i=1}^t x_{i1} \quad (4)$$

and equation (2) becomes

$$\begin{aligned} 0 < n \leq R^* & \quad f^*(n) = n \\ R^* < n \leq N & \quad f^*(n) = R^* \end{aligned} \quad (5)$$

where $R^* = g_1^* = k_1^*$ = the number of messages in rank 1 in L*; i.e., the number of relevant messages.

An examination of equations (4) and (5), (keeping in mind that x_{i1} stands for the number of messages ranked i by the L list that are ranked 1 by the Master; i.e., are relevant; and that i can equal only

1 or 2) shows that $f(n)$ is simply the (average) cumulative number of relevant messages in the first n messages. Since we have defined R^* to be the number of relevant messages, we see that we may use these terms to write the classical recall and precision ratios (RR , PR) for a given number n of messages required to be retrieved:

$$RR(n) = \frac{f(n)}{R^*}$$

$$PR(n) = \frac{f(n)}{n}$$

Under this constraint of requiring n messages to be retrieved, even the best retrieval list possible, the Master List L^* , will have the classical ratios

$$RR^*(n) = \frac{f^*(n)}{R^*}$$

$$PR^*(n) = \frac{f^*(n)}{n}$$

The measure $\mu(n)$ may thus also be looked upon as a normalized "sliding" (with n) precision and/or recall ratio, since

$$\mu(n) = \frac{f(n)}{f^*(n)} = \frac{\frac{f(n)}{R^*}}{\frac{f^*(n)}{R^*}} = \frac{RR(n)}{RR^*(n)}$$

$$= \frac{\frac{f(n)}{n}}{\frac{f^*(n)}{n}} = \frac{PR(n)}{PR^*(n)}$$

B. A Browser's Statistic

An interesting measure seems possible if, in fact, the requestor is a Browser in the sense described in Section IV. That is, he continues to search through the collection until he finds a single message which really satisfies the query. We might, therefore, picture

the value of a message as being related to the probability that that particular message will completely satisfy the query, thus ending the need for further messages. If this is the case, then one might use, as a measure of effectiveness, the expected number of messages that it takes for a requestor to be satisfied, given a particular ranked list. This number may be calculated as follows:

Let us call the master value of the i th message (according to a specific query) $V(i)$ and, as assumed above

$$V(i) = \text{probability \{ith message satisfies the query\}}$$

In addition, we make the assumption (sometimes a risky one) that these probabilities are independent of the previous messages examined. Thus, the probability that satisfaction occurs with the n th message of the ranked list $(x_1, x_2, x_3, \dots, x_n)$ becomes

$$f(n) = V(x_n) \prod_{i=1}^{n-1} (1-V(x_i))$$

where it is assumed that the messages in the list are examined in the order presented.

If the ranking is imperfect, in that it contains t sets of ties, containing e_j ($j=1, 2, \dots, t$) messages each, then the probability that satisfaction occurs at the n th message is

$$f(n) = \frac{1}{K} \sum_{k=1}^K f_k(n) \quad K = \prod_{j=1}^t (e_j!) \quad (6)$$

where

$$f_k(n) = V(x_n^k) \prod_{i=1}^{n-1} (1-V(x_i^k))$$

x_i^k = message in position i in the k th permutation of the list, these permutations involving all possible permutations of messages within tied ranks

Once $f(n)$ is obtained for a given list, the expected number of messages until satisfaction, \bar{n} , becomes

$$\bar{n} = \sum_{n=1}^N nf(n)$$

where N is the number of messages in the collection.

It is also convenient to be able to normalize this number, which may be done by calculating n for the master list. For the master list, we note that $p_i = (\text{the rank of } x_i) = i$, which leads finally to the measure J which we call the Browser's Statistic:

$$J = \frac{\sum_{n=1}^N nf(n)}{\sum_{n=1}^N nV(n) \prod_{i=1}^{n-1} (1-V(i))} \quad (7)$$

We have not tested this measure and, unfortunately, at first glance it is unclear whether it has practical value.

C. A Weighted Rank Correlation Statistic

The M-V Statistic is essentially a technique to weight only certain portions of the list. The portion up to rank R is weighted unity, the portion past rank R is weighted zero. There seems to be a more mathematically tractable device: weighting the whole list in decreasing importance from the head of the list to the end. The first function of such sort that comes to mind is a simple exponential weighting. It might be possible to adapt Kendall's τ statistic to be subject to such a weighting. (In fact, the ideal weighting should perhaps be the actual master values of the messages themselves.) This concept of weighting might be developed, at first, in a general way.

We use the formalism leading to the definition of the general correlation coefficient:

$$\tau = \frac{\sum a_{ij}b_{ij}}{\left[\sum a_{ij}^2 b_{ij}^2 \right]^{\frac{1}{2}}} \quad (8)$$

However, instead of having the scores $a_{ij}(b_{ij})$ dependent upon just pairwise comparisons within the A list (B list), let us define the B list to be the master list, so that

$$q_i = \text{rank of } i\text{th message in master list} = i$$

$$b_{ij} = \begin{cases} +1 & i < j \\ -1 & i > j \end{cases}$$

$$\sum_{i=1}^N \sum_{j=1}^N b_{ij}^2 = n(N-1)$$

Now, let us define a weighting function $w(i)$, such that $W(i)$ is a measure of the importance of having the message with master rank i being assigned rank i by the system (the A list). Then we may define

$$a_{ij} = \begin{cases} +W(i) & p_i < p_j \\ 0 & p_i = p_j \\ -W(j) & p_i > p_j \end{cases}$$

where again

$$p_i = \text{rank } i\text{th message in the A list (system list)}$$

An exponentially weighted list would then have

$$W(i) = \alpha^{i-1}$$

where α is some number between 0 and 1. The resulting coefficient (8) would be a function of α . We shall call this coefficient τ_α .

Note that if $\alpha = 1$, τ_α becomes simply Kendall's τ . When $\alpha = 0$, then τ_α becomes +1 when $p_1 = 1$, and 0 when $p_2 = 1$ (so that only the first document in the list is counted). A "master weighted" list might have $W(i) = V(i)$.

The algebraic properties of such weighted statistics have yet to be determined.

D. Cost Matrix Measures

Another way of looking at the value of a system list is as follows. Consider the matrix $\{C_{ij}\}$. This matrix represents the cost of having a message whose master ranking is i , assigned the rank j by the system. If it is possible to associate cost penalties to each such assignment independent of other assignments, then the value of a given list of assignments may be obtained from such a matrix.

This may be accomplished by constructing an "occupancy matrix" O_{ij} , where

$$O_{ij} = \begin{cases} 1 & \text{when message with master rank } i \text{ is} \\ & \text{assigned rank } j \\ 0 & \text{otherwise} \end{cases}$$

Using this matrix, one obtains for the "cost" of the entire list:

$$C = \sum_j \sum_i O_{ij} C_{ij} \quad (9)$$

The criterion matrix $\{C_{ij}\}$ may be used to extend the concepts (or our preconceptions) concerning what a retrieval system really does. One might look at a retrieval system in the following way. Given a particular query, there is a certain probability that a message will be ranked j given that its true, or master, ranking is i . Thus, a stochastic matrix, P , may be created, where P_{ij} is simply this probability. The criterion matrix may be used to evaluate any particular given list, as long as the occupancy matrix O_{ij} consists of only a single 1 in any

column or row. It is possible to conceive of some similar operation of the P matrix on this criterion matrix, such that for any given system represented by a P matrix, an over-all average effectiveness of the system might be obtained. There are some desirable elements that this operation should have. The measure obtained should be easily normalized so that:

(a) it is +1 when the P matrix is the identity matrix;

(b) it is -1 when the P matrix is the anti-diagonal matrix;

$$P_{ij} = \begin{cases} 1 & i = N - j + 1 \\ 0 & \text{otherwise} \end{cases} ;$$

(c) it is 0 when $P_{ij} = \frac{1}{N}$ for all i, j (i.e., when all lists are "randomly" created).

APPENDIX C

EXTRAPOLATION OF OVERLAP RESULTS AND ESTIMATES OF COLLECTION PARAMETERS

This appendix summarizes our best estimates of the overlap properties of our experimental vocabularies, message items, and index sets. It provides an amplification of results discussed in Section V, Subsection B.

The structure of the situation in simplified form is as follows. We are given two sets of data items that are in 1:1 correspondence. The one consists of document abstracts (messages), the other consists of the UNITERM surrogates (term sets) for the document corresponding to the abstracts. When the messages are automatically indexed, we then have two index sets in 1:1 correspondence with respect to the document item indexed. The words in these index sets overlap, as well as fail to overlap, in ways that are of distinct interest.

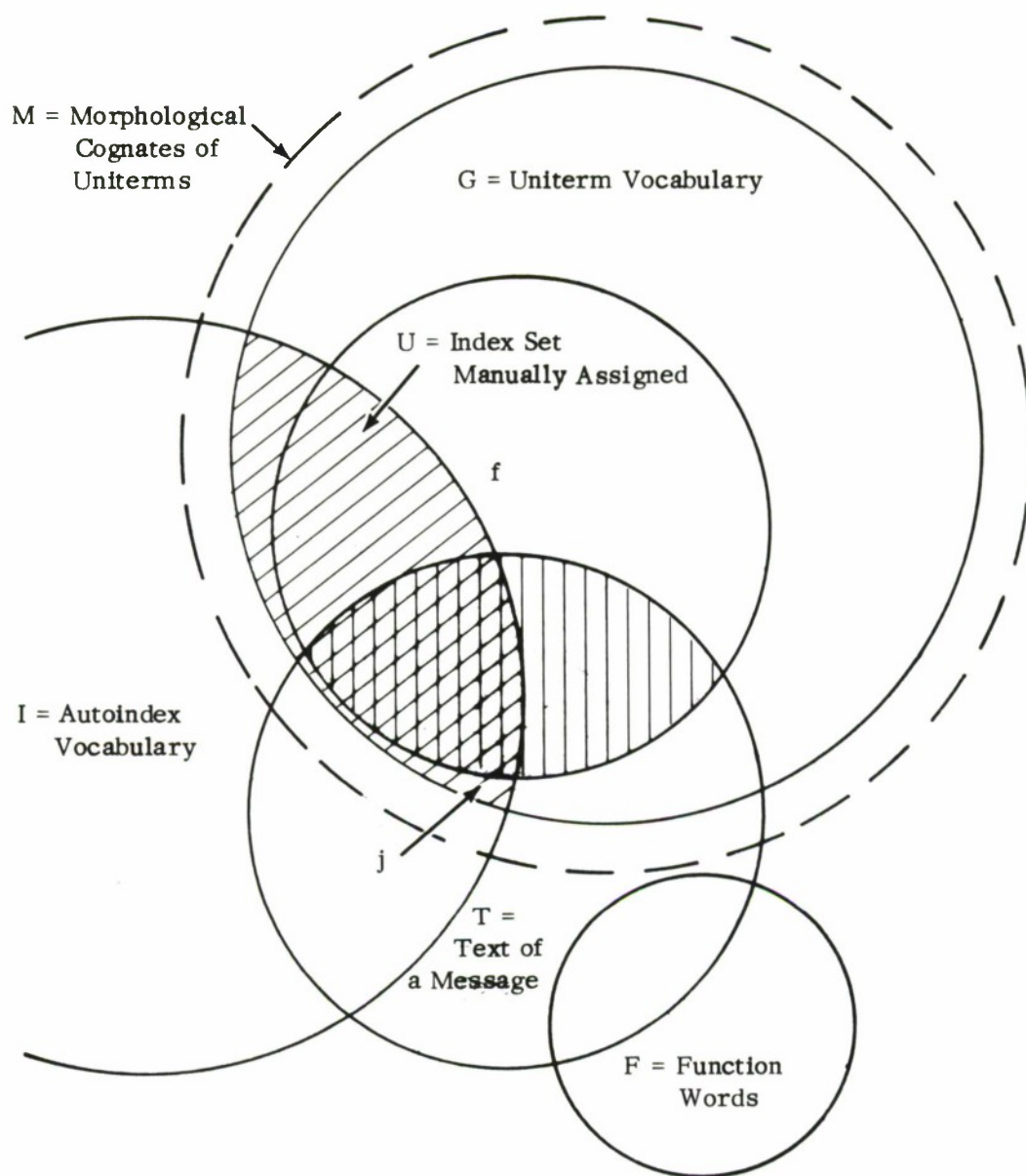
In working with these collections we have determined many important parameters -- some by exhaustive count, some by sampling, some by estimate, and some by inference. As appropriate, the details of the procedures employed have been reported elsewhere. But, with the objective of providing a unified picture of the relationships between the manual indexing of the messages and the text of the abstract, we collect in this subsection the best available description of the collections' important features.

Because a large number of sets of various kinds are involved, the situation is complicated.

To provide a framework for the presentation, and to permit each set to be explicitly named, we use the Venn diagram in Table C-1. In this table, all the sets contain types. Thus, the set I contains the 999 terms of the GE-2A vocabulary. The set M consists of all the morphological cognates of terms in G (i.e., variant forms of the UNITERM that have comparable meaning). The set F represents our exception list of 240 Function words.

A typical message is represented by the set T of the word types* that appear in it. Estimates of how many of these overlap the various vocabularies are important descriptive parameters of the collection. Similarly, overlap figures for the typical GE-OS UNITERM set surrogate (represented in the table by the term set U) provide additional descriptive information. Finally, and perhaps of greatest interest, it is valuable to extrapolate the overlap figures developed

* We systematically consider the singular and plural forms of terms in the GE-2A vocabulary to be indiscernible and always count them as one type. But we do not coalesce word forms other than those in the GE-2A vocabulary.



$U \cap T$ = Index Uniterms in Text



$I \cap G$ = Vocabulary Overlap



$T \cap I \cap U$ = Indexing Overlap: Words Assigned to a Message by both Indexing Methods

TABLE C-1 VENN DIAGRAM OF INDEXING OVERLAP

in the preceding subsection for the case where U and T both refer to the same message. In this case, note that $T \cap I$ is the set GE-2S of terms assigned by automatic indexing of the message.

a. Overlap Results

The following overlap results were either obtained by actual count or by inference from highly accurate data. These figures are considered to be very reliable.

<u>Name of Set</u>		<u>Size</u>	<u>Comment</u>
Auto Index Vocabulary	I	999	Actual count
UNITERM vocabulary	G	4824	Actual count
Function Word List	F	240	Actual count
Vocabulary Overlap	$I \cap G$	730	Actual count
Overlap Counting Cognates	$I \cap M$	880 ± 10	Estimate based on count

b. Estimates of Collection Parameters

We now turn attention to those "average" figures which have been determined for the collections as a whole.

(i) Averages for UNITERM Indexing

We consider first the collection of UNITERM sets received from GE. Were we to draw any such UNITERM set at random from the collection, i.e., to choose a random document, we would obtain the set we call U° below. As indicated, we expect to find that 9.5 of the 12.4 UNITERMS in this set will be words that are also in the Automatic Indexing vocabulary. Crudely speaking, 3/4 of the UNITERMS assigned are expected to be from the overlap vocabulary. These words would be assigned as automatic indexing terms to the chosen message if they appear in the abstract. The remaining 1/4 could not possibly be assigned by auto-indexing even if they do appear in the text of the abstract.

U^0 = the typical UNITERM set drawn from the set of 69,668 such surrogates in the GE-0 collection

<u>Name of Set</u>	<u>Size</u>	<u>Comment</u>
Surrogate U^0	12.4	Average terms per document = $\frac{862,007 \text{ postings}}{69,668 \text{ documents}}$
Overlap $U^0 \cap I$	$9.5 \pm (0.77 \pm 0.02)(12.4)$	

Notice that the figure $77 \pm 2\%$ represents the proportion of all the 862,007 postings to the 4824-term vocabulary G which are accounted for by the 730 terms in $G \cap I$.

(ii) Averages for Automatic Indexing

Since the 10,289 messages we have automatically indexed were chosen essentially at random from the set of all 45,000 abstracts which we received from GE, the parameters for that sample can reasonably be extrapolated to apply to that whole set of abstracts. Also, since the abstracts which GE did not send were presumably similar in length, degree of specificity, and other language parameters to those that were sent, we can reasonably infer that the figures below apply to the whole abstract collection.

Thus, suppose we were to choose a document at random from GE's file as we did in subsection (i) above. Looking at the abstract now, let T^0 be the set of all word types in this abstract.

<u>Name of Set</u>	<u>Size</u>	<u>Comments</u>
All types in the abstract T^0	32 (est)	44.5 word tokens (whole text)
GE-2A words $T^0 \cap I$	16.7(exact)	20.9 word tokens (GE-2A words)
Function words $T^0 \cap F$	9 (est)	14.1 word tokens (function words)
Other words	6 (est)	9.5 word tokens (other)

Only the figure for $T^0 \cap I$ is really reliable for the size of the set of types. The token figures above are accurate.

A revealing parameter for this "average" abstract is the number of words in the text that are in both indexing vocabularies. These are tags which would be assigned by auto-indexing of the message and which might also be assigned by the human indexer. Our estimate is given below, and we examine how often the indexer does assign the word in the next subsection.

<u>Name of Set</u>	<u>Size</u>	<u>Comment</u>
$T^0 \cap I \cap G$	13 (est)	$0.8(20.9) = 17$ (est) word tokens

(The 0.8 figure is based on the 85% shown in Table V-4, reduced to compensate for the exclusion of X-G terms in that development.)

c. Message-by-Message Comparison

(i) All Messages

We next consider drawing a document at random from the GE collection and examine the abstract and the UNITERM set "side by side."

When T and U refer to the same document, let

$$H = T \cap U \cap I$$

be the intersection of the two index sets assigned.

By the message-by-message comparison reported in Subsection C, it was shown that the UNITERM set assigned to the documents we studied contained, on the average, 8.3 terms which were also assigned by the auto-indexing procedure. This intersection figure on a message-by-message basis is somewhat inflated by the fact that the sample of 50 messages we used turned out to contain an inordinate number of "heavily indexed" documents; there were 20.9 UNITERMS per message vs. the expected average 12.4. Moreover, the abstracts were also somewhat "longer" than average -- with 18.2 auto index terms rather than the 16.7 average.

Taking these factors into account, we conclude that a better estimate of the average size of H for the

collection as a whole is 7, rather than the figure 8.3 obtained for the sample.

For the average document in the GE collection:

Of the 32 word types in the abstract, 7 are assigned by both indexings

$P(W \in H / W \in T) = 7/32 = 0.22$ (est) = probability that word W is assigned both by the machine and as a UNITERM, given that it is present in the text

Of the 16.7 words selected from the text of the abstract by automatic indexing, 7 are also assigned by the human indexers

$P(W \in H / W \in T \cap I) = 7/16.7 = 0.42$ (est) = probability that word W is assigned as a UNITERM, given that it is assigned by machine

Of the 13 words in the text that are listed in both indexing vocabularies, 7 are assigned to the document by both indexings.

$P(W \in H / W \in T \cap I \cap G) = 7/13 = 0.54$ (est) = probability that word W is assigned as a UNITERM, given that it is assigned by machine and that it is in the UNITERM vocabulary

Of the 12.4 UNITERMS which are assigned to the message, 9.5 are words in the automatic indexing vocabulary. Of these 9.5, 7 are actually present in the text of the given message.

$P(W \in H / W \in U \cap I) = 7/9.5 = 0.73$ (est) = probability that word W is assigned by machine, given that it is assigned as a UNITERM and that it is in the GE-2A vocabulary

(ii) Messages Without Spurious UNITERMS

We discussed in Subsection C the fact that spurious terms -- UNITERMS that did not correspond to material treated in the abstract -- were prevalent in the indexing of some messages. There exists, however, some subcollection of items where spurious terms are not a problem; i.e., where the abstract is indexed by the UNITERMS assigned. Based on available data, we can express what we would expect to observe in this "Population R" were it isolated, and that is the purpose of the development which follows. Although there is no way to check these estimates except by intellectual examination of every message in a sample, they have descriptive value and are worth reporting for that purpose.

We expect that Population R will tend to have few messages with large UNITERM sets U . (Messages with many UNITERMS are a priori more likely to carry spurious terms than "shorter" ones.) But the effect of this on the average number of UNITERMS per message will, we believe, be slight, and we expect that U will average 11-12 terms per message instead of 12.4

We would expect that the UNITERMS would be systematically more densely posted to the overlap vocabulary $G \cap I$ when we isolate the Population R. On a message-by-message basis, therefore, we expect $U \cap I$ to increase from 9.5 to at most 10. Our best estimate for the size of $U \cap I$ is 9.8.

The parameters of T will be unchanged in all respects. Also, the indexing vocabulary would be expected to be virtually the same, even if a new set of messages were used. We would still expect 16.7 terms per message, with the same bell-shaped distribution. We would expect the major change to lie in the overlap of the two index sets. The size of $U \cap T \cap I = H$ would rise from 7 to about 8 for Population R.

Based on these estimates we obtain the following descriptive parameters for the case when the abstracts and the UNITERM indexing cover the same material.

For Population R:

Of the 32 word types in the message text, 8 are assigned to the message by both indexings

$$P(W \in H / W \in T) = 8/32 = 0.25 \text{ (est)}$$

(This compares with 0.21 for the average GE document.)

Of the 11.5 UNITERMS assigned to the abstract by the indexer, 8 are also assigned by the automatic indexing procedure.

$$P(W \in H / W \in U) = 8/11.5 = 0.70 \text{ (est)}$$

(This compares with 0.56 for the average GE document.)

Of the 16.7 words selected from the text of the abstract by the automatic indexing, 8 are also assigned by the human indexer.

$$P(W \in H / W \in T \cap I) = 8/16.7 = 0.48 \text{ (est)}$$

(This compares with 0.42 for the average GE document.)

Of the 13 words in the text that are listed in both indexing vocabularies, 8 are assigned to the abstract by both indexings.

$$P(W \in H / W \in T \cap I \cap G) = 8/13 = 0.62 \text{ (est)}$$

(This compares with 0.54 for the average GE document.)

Of the 11.5 UNITERMS which are assigned to the message, 9.8 are words in the automatic indexing vocabulary. Of these 9.8 candidates, 8 are actually present in the text of the given message.

$$P(W \in H / W \in U \cap I) = 8/9.8 = 0.82 \text{ (est)}$$

(This compares with 0.73 for the average GE document.)

d. Remarks

(1) The principal disadvantage of the manual indexing from the point of view of retrieving the text of the abstracts lies in the presence of spurious terms.

(2) A strong resemblance between the automatic indexing and the manual indexing is observed for the whole collection. Selection of the more appropriate subpopulation R (i.e., those abstracts for which manual indexing does not

generate spurious term assignments) would increase that resemblance.

(3) Were the subpopulation R to be selected, it would differ little in its overlap parameters from those observed for the collection as a whole. While a tendency toward greater vocabulary overlap is present, it is only a trend: automatic indexing of the selected abstracts would still not look like a replica of the manual indexing.

(4) It would be interesting to determine if the set labeled f in the Venn diagram (UNITERMS assigned to the document which happen to be in the GE-2A vocabulary but which did not appear in the text and were, therefore, not assigned by the automatic indexing process) is peculiarly rich in terms statistically associated with those in $T \cap I$ (the set of words assigned to the message by automatic indexing). It would not be surprising to find indexers assigning associated words not present in the text.

(5) Similarly, the size of j (words which are in the UNITERM vocabulary, are present in the message and are assigned by the automatic indexing process, but which were not assigned to the document by the indexer) is indicative of disagreement between the manual and automatic procedures.

(6) Given a collection (like Population R) where there is reasonable assurance that the message and the manually-assigned tags are describing the same thing, thorough study of the size of U, T, f, j, and H would be revealing. Without suggesting that the purpose of automatic indexing is to emulate the manual indexing, we feel that the facts about the overlap are of great aid in interpreting the situation. Accurate distributional data on the sizes of the various overlapping sets is an excellent basis upon which to reason about the comparative performance attributes of retrieval systems based on either indexing.

Further observations on this last topic -- expected retrieval performance as a function of the indexing -- are discussed in conjunction with retrieval operations in Section VI.

APPENDIX D

RELIABILITY TEST OF NASA - GE OVERLAP PROPORTIONS*

A. Introduction

The overlap properties of the NASA two-word index strings and the GE-2A machine indexing vocabulary are reported in our Technical Note CACL-31. From the data displayed there, we have been able to extract some valuable estimates of the proportions of all NASA two-word index strings which fall into certain categories. These categories, as well as the appropriate proportions for both tokens and types, are displayed and discussed in Subsection A of Section VI. In Table D-1, we repeat the proportions for types for subsequent reference.

But how precise are the proportions given above? With what confidence can we state that, say, the probability that a two-word string belongs to category (b) is within the interval (0.70,0.76), or with what confidence can we state that it lies within the interval (0.66,0.80)?

This note details the development of sufficient theory to compute the error bound and determine the confidence level for each bound. It also proceeds to actually compute the error bounds for

Category	Proportion of all Two-Word Strings Types
(a) At least one of the query words is in GE-2A	0.89
(b) Second query word is in GE-2A	0.73
(c) Both query words are in GE-2A	0.30
(d) Neither query word is in GE-2A	0.11

TABLE D-1

* The material included in this appendix is due to P. Bono and S. Peters.

the type proportions given in Table D-1. Throughout the development, we make a number of assumptions, mostly conservative in nature, which are noted and discussed in detail as we proceed.

B. Notation and Theory

The data were collected from a frequency-ordered dictionary of all 18,292 NASA index terms. Consider the subset which is composed solely of two-word strings.* This subset of all two-word strings shall be called the parent population, G_T . Because the density of two-word strings is not uniform throughout the frequency-ordered dictionary, we decided to divide the parent population G_T into ten subpopulations G_1, \dots, G_{10} , and to sample each subpopulation G_i separately and independently. Let N_T be the actual number of two-word strings in G_T ** and let N_i be the actual number of two-word strings in the subpopulation G_i for $i = 1, \dots, 10$. Now, let n_i be the size of the sample, S_i , drawn from the subpopulation G_i . That is, we choose n_i two-word strings to represent the N_i two-word strings in G_i . Table 2 gives the detailed sampling procedure for each subpopulation G_i . Finally, we can define an "explosion factor" f_i which merely indicates how many strings of G_i , each string in the sample S_i represents. That is:

$$f_i = \frac{N_i}{n_i} \quad (1a)$$

This can also be written:

$$N_i = f_i n_i \quad (1b)$$

When we calculate these "explosion factors" f_i by the use of equation (1a), the reader should note that we do introduce slight error since the numbers N_i are not known with complete accuracy, but represent close estimates with their own inherent random error element. However, the error so introduced is so small, as compared with the sampling errors, that we can assume the f_i to be precise values.

Since we desire to find error bounds for each of the proportions given in Table D-1, we must also distinguish among the categories. Let

* See TN CACL-29, Figure 3, for figure showing the distribution of all Multiple Word Terms, not just two-word strings.

** This has been estimated in TN CACL-31 to consist of about 8800 terms.

Subpopu- lation	No. of Pages in Interval	How Sampled	Total Sampled
G ₁	12	Exhaustive	122
G ₂	6	First six two-word strings on each page*	36
G ₃	6	First five in each column on each page	60
G ₄	12	First three in each column on each page	72
G ₅	12	First two in each column on each page	48
G ₆	24	First one on each page	24
G ₇	24	First one on every other page	12
G ₈	24	First one on every other page	12
G ₉	24	First one on every other page	12
G ₁₀	24-1/2	First one on every other page (except none from last half-page)	12
<p>*Note: Each page consists of two columns with 54 index terms to a column.</p>			

Sampling Procedure for Subpopulations

TABLE D-2

the generic (k) represent the categories (a), (b), (c), or (d) which have been described in Table D-1.

We now proceed to choose a sample, S_i , from the subpopulation G_i (see Table D-2). To each two-word string (j_i) chosen, we can assign the random variable $x_{j_i}^{(k)}$, where

$$x_{j_i}^{(k)} = \begin{cases} 1 & \text{if the chosen two-word string } j_i \text{ belongs} \\ & \text{to category } k \\ 0 & \text{if the chosen two-word string } j_i \text{ does not} \\ & \text{belong to category } k \end{cases}$$

Since we choose a sample size of n_i , j_i runs from 1 to n_i .

Define $S_i X_i^{(k)}$ to be the actual total number of two-word strings in S_i which belong to category (k). Then

$$S_i X_i^{(k)} = \sum_{j_i=1}^{n_i} x_{j_i}^{(k)} \quad (2)$$

and $S_i X_i^{(k)}$ being a function of random variables, is also a random variable.

Finally, define $p^{(k)}$ to be the underlying probability throughout the whole parent population that a chosen two-word string belongs to category k. That is, we assume that each two-word string of the parent population has the same average probability, $p^{(k)}$, of belonging to category k.

Admittedly, this assumption disregards differences among the ten subpopulations and leads to a somewhat inflated estimate of the error bound, but a more refined model (in which, for example, we might assume separate probabilities for each of the subpopulations) would have the effect of increasing accuracy and, consequently, reducing the estimated error bound.

When we sample the subpopulation G_i , it should be noted that as soon as the first two-word string is drawn, it is no longer eligible for being chosen again and consequently the population size of G_i is reduced by one. After the next one has been drawn, it too is no longer eligible and the population is reduced by one again. This process continues

until n_i two-word strings have been chosen. Such a sampling procedure is called sampling without replacement, and the random variable, the number of chosen two-word strings belonging to category k , is hypergeometrically distributed. That is, its mean and variance are given by*:

$$E\left(\sum_{i=1}^n X_i^{(k)}\right) = n_i p^{(k)} \quad (3)$$

$$\text{Var}\left(\sum_{i=1}^n X_i^{(k)}\right) = n_i p^{(k)} q^{(k)} \left(\frac{N_i - n_i}{N_i - 1}\right) \quad (4)$$

where N_i = size of population G_i ; n_i = size of sample S_i ;

$p^{(k)}$ = pr {the chosen two-word string belongs to category k } ;

$q^{(k)}$ = pr {the chosen two-word string does not belong to category k } ;

and $p^{(k)} + q^{(k)} = 1$.

We now wish to estimate the total number of two-word strings belonging to category (k) which are in the whole subpopulation G_i .

Let $\sum_{i=1}^n X_i^{(k)}$ represent this random variable. It may be estimated by:

$$\sum_{i=1}^n X_i^{(k)} = \frac{N_i}{n_i} \sum_{i=1}^n X_i^{(k)}$$

Recalling equation (1a), we get:

$$\sum_{i=1}^n X_i^{(k)} = f_i \cdot \sum_{i=1}^n X_i^{(k)} \quad (5)$$

The expectation and variance of this random variable can be computed easily using relations (1b), (3), (4), and (5).

* The interested reader may refer to Kendall, M. G. and Stuart, A., The Advanced Theory of Statistics, Vol. 1, pp. 133-5, for a detailed derivation of these formulas.

Mean:

$$\begin{aligned}
E \left(G_i^{X(k)} \right) &= E \left(f_i \cdot S_i^{X(k)} \right) && [\text{by 5}] \\
&= f_i E \left(S_i^{X(k)} \right) \\
&= f_i n_i p^{(k)} && [\text{by 3}] \\
E \left(G_i^{X(k)} \right) &= N_i p^{(k)} && [\text{by 1b}] \quad (6)
\end{aligned}$$

Variance:

$$\begin{aligned}
\text{Var} \left(G_i^{X(k)} \right) &= \text{Var} \left(f_i \cdot S_i^{X(k)} \right) && [\text{by 5}] \\
&= f_i^2 \text{Var} \left(S_i^{X(k)} \right) \\
\text{Var} \left(G_i^{X(k)} \right) &= f_i^2 n_i p^{(k)} q^{(k)} \left(\frac{N_i - n_i}{N_i - 1} \right) && [\text{by 4}] \quad (7)
\end{aligned}$$

Consider now the parent population G_T . Let it consist of $X_T^{(k)}$ two-word strings, which belong to category (k) . That is

$$X_T^{(k)} = \sum_{i=1}^{10} G_i^{X(k)} \quad (8)$$

Recalling our definition of N_T and using equation (1b), we have

$$N_T = \sum_{i=1}^{10} N_i = \sum_{i=1}^{10} f_i n_i \quad (9)$$

Then the desired proportion $p^{(k)}$ can be estimated by:

$$\hat{p}^{(k)} = \frac{x_T^{(k)}}{N_T} \quad (10a)$$

Recalling equations (2), (5), (8), we get

$$\hat{p}^{(k)} = \frac{\sum_{i=1}^{10} G_i^{X_i^{(k)}}}{N_T} \quad (10b)$$

$$\hat{p}^{(k)} = \frac{\sum_{i=1}^{10} \sum_{j=1}^{n_i} f_i \cdot x_{ji}^{(k)}}{N_T} \quad (10c)$$

From (10c) it may be noted that $\hat{p}^{(k)}$ is a linear combination of the original random variables $x_{ji}^{(k)}$ with coefficients which add up to unity. We can easily find the expected value of $\hat{p}^{(k)}$, employing equations (3), (9), and (10b):

$$\begin{aligned} E(\hat{p}^{(k)}) &= E \left[\frac{\sum_{i=1}^{10} f_i \cdot S_i^{X_i^{(k)}}}{N_T} \right] \quad [\text{by (10b)}] \\ &= \frac{1}{N_T} E \left(\sum_{i=1}^{10} f_i \cdot S_i^{X_i^{(k)}} \right) \end{aligned}$$

$$= \frac{1}{N_T} \sum_{i=1}^{10} E \left(f_i \cdot S_i^{(k)} \right)$$

$$= \frac{1}{N_T} \sum_{i=1}^{10} f_i E \left(S_i^{(k)} \right)$$

$$= \frac{1}{N_T} \sum_{i=1}^{10} f_i n_i p^{(k)} \quad \text{[by (3)]}$$

$$= \left(\frac{1}{N_T} \right) \left(p^{(k)} \right) \left(\sum_{i=1}^{10} f_i n_i \right)$$

$$= \left(\frac{1}{N_T} \right) \left(p^{(k)} \right) \left(N_T \right) \quad \text{[by (9)]}$$

$$E \left(\hat{p}^{(k)} \right) = p^{(k)}$$

This shows that the estimate of $\hat{p}^{(k)}$ is unbiased.

We are particularly interested in the variance since it will give us a measure of how close we can expect the value of $p^{(k)}$ to be to the true mean, $p^{(k)}$. We use equations (4), (9), and (10b) to get

$$\text{Var} (\hat{p}^{(k)}) = \text{Var} \left(\frac{\sum_{i=1}^{10} f_i \cdot S_i^{X_i^{(k)}}}{N_T} \right) \quad [\text{by (10b)}]$$

$$= \frac{1}{N_T^2} \text{Var} \left(\sum_{i=1}^{10} f_i \cdot S_i^{X_i^{(k)}} \right)$$

$$= \frac{1}{N_T^2} \sum_{i=1}^{10} \text{Var} (f_i \cdot S_i^{X_i^{(k)}}) \quad \left[\begin{array}{l} \text{since } S_i^{X_i^{(k)}} \\ \text{are mutually} \\ \text{independent} \end{array} \right]$$

$$= \frac{1}{N_T^2} \sum_{i=1}^{10} f_i^2 \text{Var} (S_i^{X_i^{(k)}})$$

$$= \frac{1}{N_T^2} \sum_{i=1}^{10} \frac{f_i^2 n_i (N_i - n_i)}{(N_i - 1)} p^{(k)} q^{(k)} \quad [\text{by (4)}]$$

$$\begin{aligned} \left(\sigma^{(k)} \right)^2 = \text{Var} (\hat{p}^{(k)}) &= \frac{\sum_{i=1}^{10} f_i^2 n_i \frac{(N_i - n_i)}{(N_i - 1)}}{\left(\sum_{i=1}^{10} f_i n_i \right)^2} p^{(k)} q^{(k)} \quad [\text{by (9)}] \quad (11) \end{aligned}$$

Since the standard deviation equals the square root of the variance, we get

$$\sigma(k) = \sqrt{\frac{\sum_{i=1}^{10} f_i^2 n_i \left(\frac{N_i - n_i}{N_i - 1} \right)}{\left(\sum_{i=1}^{10} f_i n_i \right)^2} p^{(k)} q^{(k)}} \quad (12a)$$

Thus we have derived an error bound $\sigma(k)$, which we can use to establish how reliable our estimated proportions are. From a table of the normal distribution,* we know that we can expect, with 68% probability, that the actual proportion lies within \pm one standard deviation of the estimated proportion; and, with 95% probability, we can expect the actual proportion to be within \pm two standard deviations of the estimated proportion. This can be summed up as shown in Table D-3.

Confidence Level	True Proportion is Within
0.68	$\hat{p}^{(k)} \pm 1 \cdot \sigma(k)$
0.95	$\hat{p}^{(k)} \pm 2 \cdot \sigma(k)$

TABLE D-3

C. Computation of Estimated Standard Deviation

We shall use equation (12a) to compute $\sigma(k)$ for $(k) = (a)$, (b) , (c) , and (d) ; but we shall write (12a) in a more useful form:

* Since by (10c), $p^{(k)}$ is a linear compound of 410 independent random variables, it may be assumed, by the Central Limit Theorem, to be approximately normally distributed.

$$\sigma^{(k)} = \sqrt{\frac{\sum_{i=1}^{10} f_i^2 n_i \left(\frac{N_i - n_i}{N_i - 1} \right)}{\left(\sum_{i=1}^{10} f_i n_i \right)^2} p^{(k)} q^{(k)}}$$

$$\sigma^{(k)} = \sqrt{\mu p^{(k)} q^{(k)}} \quad (12b)$$

We note that μ is a function of N_i and n_i alone (since $f_i = N_i/n_i$) and does not depend upon what category we are considering. Consequently, μ has to be calculated only once. In order to calculate μ , all we need are the 20 values N_i and n_i for $i = 1, \dots, 10$.

To get the necessary data for the computation of μ , we must turn to Tables B-I and B-II of CACL-31:

- (i) Column 3 of B-II contains $N_i = f_i \cdot n_i$;
- (ii) Column 3 of B-I contains n_i .

Table D-4 contains all of the vital information as well as intermediary calculations. As can be seen,

$$(i) \quad \sum_{i=1}^{10} f_i n_i = 8806$$

$$(ii) \quad \sum_{i=1}^{10} f_i^2 n_i \left(\frac{N_i - n_i}{N_i - 1} \right) = 712,598$$

i	n_i	f_i	f_i^2	$N_i = f_i n_i$	$n_i (N_i - n_i)$	$\frac{n_i (N_i - n_i)}{(N_i - 1)}$	$\frac{f_i^2 n_i (N_i - n_i)}{(N_i - 1)}$
1	122	1.0	1	122	0	0.0	0
2	36	6.5	42	234	7128	30.6	1285
3	60	4.7	22	282	13320	47.4	1043
4	72	9.1	84	658	42192	64.2	5393
5	48	13.8	189	660	29376	44.6	8429
6	24	58.3	3402	1400	33024	23.6	80287
7	12	116.7	13612	1400	16656	11.9	161983
8	12	116.7	13612	1400	16656	11.9	161983
9	12	116.7	13612	1400	16656	11.9	161983
10	12	104.2	10851	1250	14856	12.0	130212
Selected Totals =				8806			712598

TABLE D-4

Consequently,

$$\begin{aligned}
 & \sum_{i=1}^{10} \frac{f_i^2 n_i (N_i - n_i)}{(N_i - 1)} \\
 &= \frac{\sum_{i=1}^{10} \frac{f_i^2 n_i (N_i - n_i)}{(N_i - 1)}}{\left(\sum_{i=1}^{10} f_i n_i \right)^2} \\
 &= \frac{712,598}{(8806)^2}
 \end{aligned}$$

$$\therefore \mu = .009189$$

Recalling

$$\sigma^{(k)} = \sqrt{\mu p^{(k)} q^{(k)}} \quad (12b)$$

we can easily obtain $\sigma^{(k)}$. The necessary information as well as the final calculations are shown in Table D-5, where the $p^{(k)}$ have been taken from Table D-1.

(k)	μ	$p^{(k)}$	$q^{(k)}$	$\mu p^{(k)} q^{(k)}$	$\sigma^{(k)} = \sqrt{\mu p^{(k)} q^{(k)}}$
(a)	.009189	0.89	0.11	.000900	± 0.030
(b)	.009189	0.73	0.27	.001811	± 0.042
(c)	.009189	0.30	0.70	.001930	± 0.044
(d)	.009189	0.11	0.89	.000900	± 0.030

TABLE D-5

We see that the estimated standard standard deviations are between 0.030 and 0.045. That is, for example, for category (a), we can expect with 0.68 probability that the "true" proportion lies somewhere between 0.86 and 0.92; and that with 0.95 probability, it lies between 0.83 and 0.95.

D. Summary

Category	Confidence Level = 0.68	Confidence Level = 0.95
(a)	0.89 ± 0.030	0.89 ± 0.060
(b)	0.73 ± 0.042	0.73 ± 0.084
(c)	0.30 ± 0.044	0.30 ± 0.088
(d)	0.11 ± 0.030	0.11 ± 0.060

TABLE D-6

We exhibit Table D-6, which summarizes the main results of this note.

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Arthur D. Little, Inc. Cambridge, Mass.		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP N/A	
3. REPORT TITLE STUDY AND TEST OF A METHODOLOGY FOR LABORATORY EVALUATION OF MESSAGE RETRIEVAL SYSTEMS			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Interim Report			
5. AUTHOR(S) (Last name, first name, initial) Giuliano, Vincent E. Jones, Paul E.			
6. REPORT DATE August 1966		7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
8a. CONTRACT OR GRANT NO. AF 19(628)-4067		9a. ORIGINATOR'S REPORT NUMBER(S) C-66257	
b. PROJECT NO.		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
c.		ESD-TR-66-405	
d.			
10. AVAILABILITY/LIMITATION NOTICES Distribution of this report is unlimited.			
11. SUPPLEMENTARY NOTES None		12. SPONSORING MILITARY ACTIVITY Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, USAF, L. G. Hanscom Field, Bedford, Mass. 01730	
13. ABSTRACT This report documents two years of work on the laboratory evaluation of message and document retrieval systems. It contains a general discussion of the problems of laboratory evaluation of retrieval systems, and specific findings relating both to the methodology of evaluation and search performance results observed with a large-scale experimental collection. The initial sections of the report are devoted to developing a general framework for viewing the problems of performance evaluation under laboratory conditions. We identify and discuss several mathematical techniques potentially useful in the evaluation process, including methods for unbiased and averaging the results of judgments by several independent evaluators. Also, many possible measures of system performance are discussed, compared, and evaluated. We describe the processing of our 10,000-message experimental collection, including the steps of automatic indexing and computation of word-association measures. Comparison of subject matter coverage and effects of manual and automatic indexing for this collection are discussed, and several statistical characterizations of our collection are presented. We also describe several experimental forays with our collection using combinations of conventional and associative retrieval with and without human intervention, using multiple evaluators, and we consider both full text and subject heading queries. Numerous conclusions and findings are presented with respect to efficacy of various retrieval evaluation techniques and methods, the relative merits of machine and automatic indexing, and the comparative efficacy of various combinations of conventional and associative search options.			

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Computers						
Information retrieval						
English language						
Evaluation						
Indexing						
Automatic indexing						
Vocabulary						
Documentation						
Message						
Performance characteristic curves						
Graphical methods						
Statistics						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those

imposed by security classification, using standard statements such as:

- (1) "Qualified requesters may obtain copies of this report from DDC."
- (2) "Foreign announcement and dissemination of this report by DDC is not authorized."
- (3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."
- (4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."
- (5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.